IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 29, 2018, accepted November 13, 2018, date of publication November 19, 2018, date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2882244

Interactive Data Exploration of Distributed Raw Files: A Systematic Mapping Study

ALEJANDRO ALVAREZ-AYLLON[1], MANUEL PALOMO-DUARTE[2], AND JUAN-MANUEL DODERO[2]

[1]Geneva Observatory, University of Geneva, 1211 Geneva, Switzerland
[2]Department of Computer Science and Engineering, University of Cádiz, 11001 Cádiz, Spain

Corresponding author: Alejandro Alvarez-Ayllon (alejandro.alvarezayllon@unige.ch)

This work has been developed in the VISAIGLE project, funded by the Spanish National Research Agency (AEI) with ERDF funds under grant ref. TIN2017-85797-R.

ABSTRACT When exploring big amounts of data without a clear target, providing an interactive experience becomes really difficult, since this tentative inspection usually defeats any early decision on data structures or indexing strategies. This is also true in the physics domain, specifically in high-energy physics, where the huge volume of data generated by the detectors are normally explored via C++ code using batch processing, which introduces a considerable latency. An interactive tool, when integrated into the existing data management systems, can add a great value to the usability of these platforms. Here, we intend to review the current state-of-the-art of interactive data exploration, aiming at satisfying three requirements: access to raw data files, stored in a distributed environment, and with a reasonably low latency. This paper follows the guidelines for systematic mapping studies, which is well suited for gathering and classifying available studies. We summarize the results after classifying the 242 papers that passed our inclusion criteria. While there are many proposed solutions that tackle the problem in different manners, there is little evidence available about their implementation in practice. Almost all of the solutions found by this paper cover a subset of our requirements, with only one partially satisfying the three. The solutions for data exploration abound. It is an active research area and, considering the continuous growth of data volume and variety, is only to become harder. There is a niche for research on a solution that covers our requirements, and the required building blocks are there.

INDEX TERMS Big data applications, data analysis, data engineering, data exploration, database systems, interactive systems, systematic mapping study.

I. INTRODUCTION

Extracting knowledge from raw data is a well-known problem for many and very diverse domains—from finance to science. This is known as *Knowledge Discovery in Databases* (KDD), because ''knowledge'' is the final product of the process [1], [2]. *Data Mining* is often used as a synonym, although some authors consider it to be part of the KDD process itself rather than completely equivalent [2], [3]. We tend to agree more with the second view but this does not affect the purpose, scope or results of this study.

To help understand the scope of the current study, we refer to the CRISP-DM (CRoss Industry Standard Process for Data Mining) [4], which proposes a process model for data mining projects. The phases of this process can be seen in figure 1.

The scope of Interactive Data Exploration (IDE) tools lies on the *data understanding* phase. This has human intuition as a core part of the process, where the user tentatively explores the data, iterating and reformulating the queries as their knowledge and insight changes with each iteration. The target of this stage is to generate new hypotheses and not to validate them [5]. The validation is left for the *Evaluation* phase.

A system that is able to be used in such a way needs to be lightweight, adaptive and have reasonably low response times—[6] considers two seconds to be the upper limit for the continuity of thoughts—, helping and assisting, without getting in the way of the person involved in the loop.

These restrictions, combined with the ''data deluge'', impact almost all scientific research domains and they pose a hard and interesting problem. On the one hand, we need responsive and efficient systems for querying huge volumes of data. On the other hand, since the access patterns are

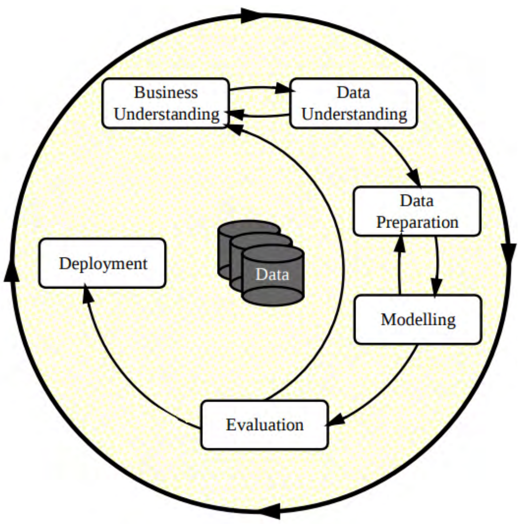

FIGURE 1. The CRISP-DM process model, from [4].

not only unknown beforehand but also variable with time, traditional approaches that enforce an early decision on data structure, storage and indexing are unsuitable [7].

This problem can be tackled at different levels—from the physical layout on disk, to the interface interacting with the user. In 2015, Idreos *et al.* [8] classified several of these solutions depending on which approach they take on the issue. This paper attracted our attention to this research area due to the potential applications in High Energy Physics (HEP) and, in particular, for the processing of ROOT files containing data from the Large Hadron Collider (*LHC*) at CERN.

This possibility is, in fact, mentioned as a motivating example of some of the papers to which we initially had access [9]–[11], although, to the best of our knowledge, they have not been implemented in practice.

For such a system to be practical, it also has to be able to run on multiple ROOT files that are distributed across several machines—located in two separate sites—at the CERN data center.

In summary, we need to satisfy three main requirements:

1) *Interactive response times*, as already discussed
2) *Access to raw data files*. Pre-loading data in main memory is not an option due to the data volume and because we aim for a system that extends and does not replace the existing data management solution
3) *Distributed*, since files are stored and replicated by an already existing distributed storage.

Ideally, the granularity of the access has to be higher than "file level" because scientists normally worry about datasets that are defined by the data origin, year, conditions, etc..., and one dataset may be distributed across several files [12].

To follow up on this idea and to identify if there is any existing solution, we have done a systematic mapping study to get a rigorous picture of the state of the art, how it has changed since Idreos' tutorial, and to determine the maturity of the area. Even thought our motivation example emerges from the HEP domain, this study is focused on interactive data exploration in general, and can be of interest for researchers in other scientific domains.

The rest of this document is structured as follows. SubsectionI-A is an overview of the different approaches that attack each of our three requirements for interactive data exploration. Section II describes how this study has been done and Section III summarizes the findings. Section IV includes a discussion of the results, including, for completeness and fairness, threats to the validity of this secondary study. Finally, Section VI lists the conclusions.

A. OVERVIEW

While in this study we do a systematic mapping study of the interactive data exploration research area in general, we were initially motivated by the three constrains for our use case: exploration of raw data files, located on a distributed storage, and with a latency low enough as to enable interactive use.

Here, we summarize some of the approaches we have found used to cover each one of our requirements.

1) RAW DATA FILES

We have to provide access to data stored in the form of ROOT files, that has a volume of several Petabytes, and which keeps growing each year [13]. While these files can be stored on tape or disk, we focus only on those available on disk, as the latency of tape storage is way beyond the interactivity requirements. Depending on the experiment, the number of files stored on disk can range between 260M to 500M [14], normally on the order of one to ten GiB [15]. This basically discards a scenario where the data is pre-loaded in main memory because it would take a considerable amount of time and, at the very least, duplicate the amount of required storage. Furthermore, the fact that the best schema design, if any, can be unknown at first makes this more difficult because it becomes completely impractical to re-design and re-load the data several times as the exploration progresses.

For these reasons, we are interested in engines that allow *in situ* queries, as proposed by [16]. In this paper, Idreos *et al.* lead a line of research that is focused on systems that are capable of executing queries over flat raw files without any preprocessing, adapting their internal working dynamically to the workflow. More specifically, they prototype an *adaptive* loading system that reads data when needed and suggests possible directions for further research on adaptive systems: storage, execution, and auto-tuning.

Following on the vision of that paper, [17] presents the "NoDB" paradigm, which provides access to raw data files avoiding the latency and overhead introduced by pre-loading, and which are comparable in performance with traditional Database Management Systems (*DBMS*). Since there is no pre-loading, data has to be read as needed—adaptive loading. The system also needs to generate indexes dynamically to remain performant.

Going one step further, [9] introduces RAW, which is a query engine capable of querying not only CSV files but also more complex files as ROOT files. This engine is based on code-generation and it uses plug-ins for specific file formats. Similarly, Proteus [10] also uses code generation to support heterogeneous data formats, traversing the query plan only once to generate the code to be compiled and executed on the fly.

With SCANRAW [18], improvements for this kind of solution are proposed by parallelizing parts of the processing and loading the data into a database system to improve the execution time of following queries.

Both Alpine [19] and Slalom [20] support queries over raw data files. They improve the adaptive indexing of raw files by also creating adaptive partitioning over the original file and deciding the most suitable indexing strategy to use separately for each partition.

In summary, for querying raw data files in a binary format, systems need to provide a plug-in mechanism that extends the original implementation with different data formats. Code generation can be used to remove the overhead caused by indirections. Given that the original files are not usually indexed, these systems also need to create assisting data structures on-the-fly to avoid the initial load time that more traditional database systems normally require.

2) INTERACTIVE RESPONSE TIMES

With large data volumes, response times can be much higher than the interactive limit of two seconds, even with good indexes. When the data is being tentatively explored, a fast ''good enough'' response can be better than a complete but much slower one.

Approximate Query Processing (AQP) [21]–[23] approaches can help when we can compromise some accuracy for better response times, reducing the amount of data to be processed for each query.

The most common and obvious approach to reduce the amount of data to be processed is *sampling*, which limits the processing to a subset of the original data. However, this introduces an associated error with any given query, which in itself is also the subject of research.

Errors caused by sampling can affect the performance of the system itself [22] and the decisions taken by the end users [22], [24] because they may be more used to the complete output provide by traditional DBMS or they may misinterpret the error estimations given by the system.

Error estimation techniques are normally classified into two main sets [22], [25], [26]:

- Analytical: These methods can be fast but they need to be manually derived for each type of query. Consequently, they are normally available only for simple queries with basic data aggregations.
- Bootstrap [27]: These are more flexible because they use re-sampling of the original sample to estimate the error. However, this makes them also more computationally expensive.

The *analytical bootstrap method* [25], reduces the overhead of the bootstrap error estimation, removing the need for re-sampling.

It is worth mentioning that sampling tends to fail when the query interest is focused on extreme values (outliers) [22], [28], [29].

Another recent approach is *database learning* [30], which exploits the answers to past queries to infer some knowledge about the nature of the underlying data, decreasing with time the amount of data to be read. Following this idea further, *active* database learning [30]–[32] proposes systems that would pro-actively ''train themselves'' to improve their models [33]. However, as of the time of this writing, we are unaware of any database system that implements this technique.

3) DISTRIBUTED ENVIRONMENT

Seaweed [34] deserves a mention for this requirement because it is the only system found by this study that clearly states its objective of scaling to a big number of end-systems (10^3 to 10^9), where it is usual to have some of them off-line or going off-line at any given moment.

These authors also consider that centralization, redistribution and replication of the data can limit the scalability of the system, especially due to the requirements imposed on the network when it has to be moved away from where it was originated.

We are interested in systems that could sit on top of an existing data storage solution where replication and distribution policies are out of our control. Thus, similarly to Seaweed, we need to process the data wherever it is located. This location may be off-line.

They solve this issue persisting the queries for a given delay, so when a back-end system comes back online it will execute its part of the plan, updating incrementally the results. This delay enables the user to reach a compromise between the completeness of the response and the responsiveness. We find that this approach can be interesting for our use case.

II. METHOD

A systematic mapping study is a process for the exploration of the situation of a wide research area with a high level of granularity, allowing us to identify areas in the domain where it may be interesting to explore in more detail [35]. Because we are trying to obtain an overview of the situation of the research on data exploration techniques and identify where additional work may be required, we have decided to follow this approach, and, more specifically, the guidelines proposed by [36]. For completeness, we include in figure 2 the diagram of the process for a systematic mapping study, as defined by Petersen *et al*.

We first justify the need for the study and the research questions it will answer. Then, we perform the search for papers from different sources, applying a selection criteria to discard those that are not of interest for the purposes of this study. We define the classification schema used to map the

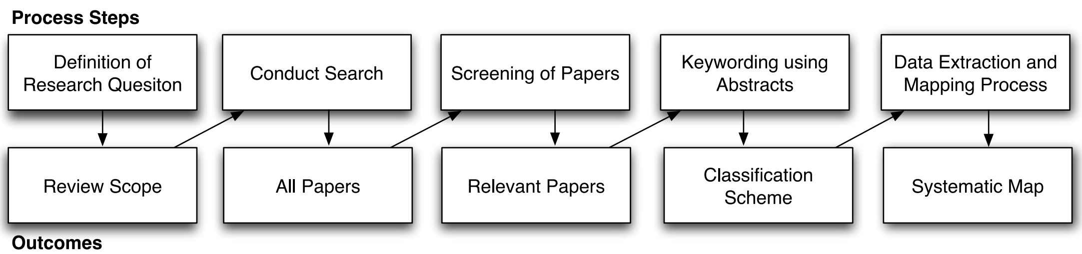

FIGURE 2. **The Systematic mapping process [36].**

current status of the domain, and, finally, we propose how to summarize and visualize the resulting data.

A. JUSTIFICATION

LHC data are stored as ROOT [37] files. Some of the analysis on these files are relatively simple queries, which is currently done with hand-written C++ programs. Even though Karpathiotakis *et al*. has already proposed using declarative queries instead [9], we are unaware of any progress in that direction since it was used as a motivating example.

A tool that provides a high level of querying this type of data can be of great use, especially if integrated with the existing storage solutions used today by the LHC experiments, such as EOS [38]. This interface would allow scientists to spend more time exploring the data and less time writing low level code to dive through the specifics of the file format.

However, before embarking on such a project, we need to get a better picture of the state-of-the-art because more recent developments may already cover part, if not all, of our needs. Systematic mapping can be a suitable tool for this purpose. Furthermore, the output of this study can help other researchers to identify interesting directions for their own work or even tools for those looking to cover a similar need.

B. RESEARCH QUESTIONS

In the tutorial, Idreos *et al*. [8] propose a classification of different possible approaches to our problem. This study provides an excellent introduction but we wanted to expand on it by answering two questions that were not covered by the original paper and we also wished to survey the subsequent evolution of the domain.

1) RQ1. HOW HAS THE RESEARCH AREA EVOLVED?

Given that this is an active research area, it has probably progressed since the Idreos *et al*. tutorial that we are using as a baseline. Therefore, the first question to answer to decide how to focus future research is: How has it evolved since 2015?

2) RQ2. WHAT IS THE MATURITY LEVEL OF THE RESEARCH AREA?

How many complete and reliable solutions are there? Are they successfully implemented in practice? How do they improve the users' experience? Identifying publications is not enough, we also want to assess in what part of the software lifecycle they focus.

3) RQ3. HOW FAR ARE WE FROM A TOOL THAT SOLVES OUR THREE REQUIREMENTS?

The final target of this research is to identify solutions that cover our three requirements and could be integrated into the storage software at CERN. Even though Idreos *et al*. [8] closed their tutorial by mentioning the importance of interconnection research, they do not provide any references or study on this area.

C. SEARCH STRATEGY

For the retrieval of studies, it is necessary to clearly define how the search is going to be performed. This work combines three different strategies, as follows:

- Set of known works obtained from [8] because our RQ2 is not covered by the original classification.
- Forward snowballing [39] from the known set of publications using Google Scholar.
- For completeness, database searches to improve the coverage of our study.

Jalali and Wohlin [40] argue that snowballing and database searches can lead to similar patterns but they also agree that it is "not easy to draw any general conclusions" about if the conclusions obtained are the same using the two different approaches. Thus, we have opted to follow both.

The set of digital libraries consulted is:

- ACM Digital Library
- Elsevier (Science Direct)
- Springer
- IEEE Digital Library
- Wiley Online Library
- World Scientific Net

Given the fast pace at which the field moves, older papers have been probably superseded or, if still relevant, we expect them to be already included in [8]. Consequently, we have limited the scope in time to studies published from 2010 onwards

All of the references obtained by any of the previous method were imported into a group in the *Mendeley Reference Manager*. Any obviously non-interesting entry — such as book or proceeding indexes—were removed at this

TABLE 1. Category.

User Interaction			
Data Visualization	Visual Optimizations	Visual Tools	
Exploration Interfaces	Automatic Exploration	Assisted Query Formulation	Novel Query Interfaces
Middleware			
Interactive Performance Optimizations	Data Prefetching	Query Approximation	
Database Layer			
Indexes	Adaptive Indexing	Time Series	Flexible Engines
Data Storage	Adaptive Loading	Adaptive Storage	Sampling

stage. The definitive list can be found on a public group in Mendeley.com[†]

D. STUDY SELECTION CRITERIA

We based the initial screening of studies on title, abstract, and keywords. In some cases, when the information provided by these fields was insufficient to take a decision, we also considered their conclusions or read the complete study.

We have focused here on finding primary studies related to data exploration. The filtering was performed using the following exclusion criteria:

Unsupported language	Studies written in a language different than English, Spanish or French
Incomplete publication	Abstract only, or presentations were excluded
Off topic	Out of the data exploration domain
Not a primary study	Secondary, tertiary and surveys
Duplication	In case of duplication, or high similarity for the same set of authors, only the most complete or the most recent was taken into account.

Those publications that passed the inclusion criteria were reviewed to make sure all their fields were correct. Normally, this should have been done during the previous stage but due to the sheer volume of publications yielded by the search strategy this step was postponed until the filtering was done. Because only title and abstract were used for the filtering, this did not affect the end result.

E. CLASSIFICATION

Publications that pass the selection criteria will be classified into two axes: data exploration facet and research type.

1) CATEGORY

As mentioned in section II-B, we base our study on the classification done by Idreos *et al.* [8], which is included for convenience in table 1. For more details, we refer the interested reader to Idreos' tutorial.

For our purposes, we have assigned one single category to each work covered by our study, choosing the most prominent topic when more than one category could fit.

[†] https://www.mendeley.com/community/interactive-data-exploration-in-science-systematic-mapping/

TABLE 2. Research type.

Research type	Description
Evaluation research	Investigation of a problem or implementation in practice.
Proposal of solution	These papers propose a solution and argue for its relevance without complete validation. A proof-of-concept may be offered.
Validation research	These papers investigate the properties of a solution proposal that has not yet been implemented in practice.
Philosophical papers	These papers sketch a new way of looking at things, a conceptual framework, etc.
Opinion papers	These paper contain the author's opinion.
Personal experience papers	These paper should contain a list of lessons learned by the author from his or her own experience. The evidence can be anecdotal.

2) RESEARCH TYPE

To answer our second research question—the maturity of the area—we follow the classification of research approaches done by [41], as our guidelines for systematic mapping do [36].

We summarize the different research types in table 2.

As per this classification, we expect mature solutions that have been implemented in practice to be covered by one or more *Evaluation Research* studies. If, on the contrary, they are on very early stages, then most related studies will fall into the *Philosophical* or *Opinion* categories.

F. DATA EXTRACTION AND VISUALIZATION

At this stage, the papers were filtered and classified. We needed to summarize the obtained data in a way that is useful to answer our research questions.

To answer *RQ1*, we focused on the counting of each category and their visualization on a time series plot.

To answer *RQ2*, a bubble plot can help to more easily identify the most frequent research type per category. In this way, we can identify if one area is more mature than other. Additionally, we also counted and displayed how many publications include some sort of user study, which should prove if any particular solution is successful at improving the integration of a human on the loop.

Finally, for *RQ3*, we flag interesting papers classified under *Proposal of Solution* with the three requirements separately, if stated on their abstract or conclusions.

Additionally, while it was not in the original research questions, we can also extract which publication forums are the most prominent on our results.

TABLE 3. Search queries.

Library	Scope	Search
ACM Digital Library	Full text	("RAW data" OR "RAW file" OR "ROOT file") AND (query OR exploration)
ScienceDirect	Title, abstract, keywords (computer science)	((RAW OR ROOT) AND (query OR exploration))
Springer	Full text (computer science)	("RAW data") AND (query OR exploration) + ("RAW file") AND (query OR exploration)
Wiley Online Library	Abstract	RAW AND query
IEEE Digital Library	Abstract	RAW AND query
World Scientific Net	Full text (computer science)	RAW AND query

TABLE 4. Accepted and rejected count.

Accepted	Duplicated	Not Primary	Off Topic	Too Old	Total
242	9	16	5,295	126	5,688
4.25%	0.16%	0.28 %	93.09%	2.22%	100%

III. RESULTS

In this section, we describe the outcome of each stage of the systematic mapping.

A. STUDY SELECTION

As previously described, we have three different sources of papers: the references from [8], search engines, and forward snowballing from those that pass the selection criteria.

Table 3 displays the search queries that were used for each digital library. All searches were done on May 16, 2017 and they yielded a total of 5,525 articles.

Idreos' tutorial provided 47 papers and the forward snowballing provided 116.

From this total of 5,688, only 242—4.25%–were accepted, the details are shown in table 4. This rather low hit ratio comes mostly from the on-line searching of digital libraries because the lack of well defined, or univocal, keywords makes it difficult to decide what to search for. We do not seem to be alone in this respect [42], [43].

Even once defined, and because we must use different search engines, there are few or no commonalities between the way queries can be written and handled between different archives [44], [45].

This yield is no smaller than those of systematic studies in other fields, which can be as low as 0.3% [46].

B. STUDY DATA EXTRACTION

Table 5 displays the frequency of publications for each classification cluster proposed by [8]. It is worth mentioning that four papers on the *Database Layer* did not fall into the predefined clusters, given their genericity [7], or as an evaluation of different techniques [47]–[49].

Figure 3 displays the frequency of each major cluster against the research type count for each one. In table 6, we display the publication forums where more than one study has been published. While there are two main forum, summing 30.58% of all the publications, most of the papers are spread out on different conferences and journals.

It is worth noting that this table includes gray literature; that is, outside of the formal academic publishing. While one may argue that this papers have not been [yet] subject of a peer review, they are still included because gray literature can be, and is, a useful source of knowledge for information users [50]. In fact, Kitchenham and Charters [35] recommended in their guidelines for systematic reviews to include gray literature in searches.

TABLE 5. Category summary.

Category	Count
User Interaction	**86**
Assisted Query Formulation	28
Visual Optimizations	25
Novel Query Interfaces	14
Visualization Tools	11
Automatic Exploration	7
Exploration Interfaces	1
Middleware	**48**
Query Approximation	34
Data Prefetching	14
Database Layer	**108**
Adaptive Indexing	26
Flexible Engines	16
Time Series	16
Sampling	15
Adaptive Storage	14
Adaptive Loading	10
Spatial Query	6
Other	5

IV. DISCUSSION

A. ANSWERING THE RESEARCH QUESTIONS

1) RQ1. HOW HAS THE RESEARCH AREA EVOLVED?

Figure 4 displays the evolution during time of each of the three major classification clusters: user interaction, middleware and database.

Considering our search strategy, most of the results are posterior to 2012. Different approaches seem to be, in general, well balanced—we refer again to table 5—, although there is space for more works focused on *exploration interfaces* and *automatic exploration*, which are the less frequent published approaches.

2) RQ2. WHAT IS THE MATURITY LEVEL OF THE EXISTING SOLUTIONS?

We can use the figure 3 to answer this question. The vast majority of papers considered by this study—79.35%—fall within the *proposal of solution* research type.

Meanwhile, *evaluation* and *validation* research are represented just by a 11% and 6.07%, respectively. Only 32 documents (13%) include some sort of user study:

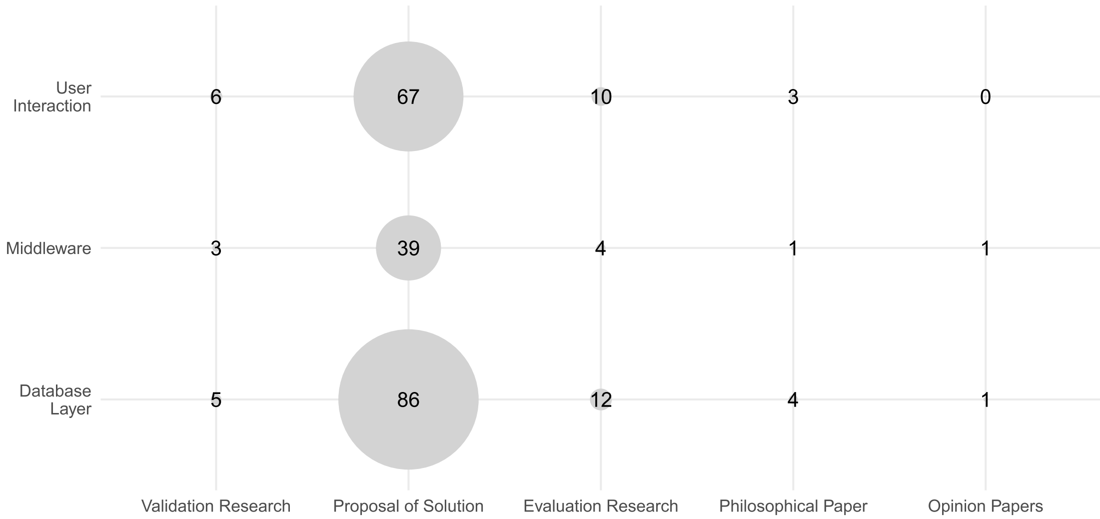

FIGURE 3. Layer vs research type.

TABLE 6. Publication forum.

Publication	Count
Journal	**55**
The VLDB Journal	11
IEEE Transactions on Knowledge and Data Engineering	3
IEEE Transactions on Visualization and Computer Graphics	3
International Journal of Cooperative Information Systems	3
Journal of Big Data	3
ACM Transactions on Database Systems	2
Future Generation Computer Systems	2
SIGMOD Record	2
Others	26
Conference	**181**
ACM International Conference on Management of Data (SIGMOD)	33
Proceedings of the VLDB Endowment	30
IEEE International Conference on Data Engineering	11
Conference on Innovative Data Systems Research (CIDR)	9
Database Systems for Advanced Applications	5
International Conference on Scientific and Statistical Database Management	5
IEEE International Conference on Big Data	4
International Conference on Extending Database Technology	3
International Workshop on Data Management on New Hardware	3
ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)	2
Advances in Visual Computing	2
Big Data Analytics	2
Database and Expert Systems Applications	2
IEEE International Conference on Mobile Data Management	2
Intelligent Information and Database Systems	2
International Conference on Advanced Cloud and Big Data	2
Workshop on Human-In-the-Loop Data Analytics	2
Others	62
Gray literature	**6**

24 for ‘User Interaction’, 4 for ‘Database Layer’ and 2 for ‘Middleware’. Research on how different solutions —either existing or proposed— perform in practice is lacking.

These figures are hardly surprising because they seem to have been commonplace in computer science for a long time now [51]–[53]. For instance, Sjöbergh *et al.* [53] survey the status of controlled experiments in software engineering and the numbers they find are equally low, with only 113 controlled experiments found on 5,453 papers.

It is hard and also out of the scope of this study to make some inferences from these results. Tichy *et al.* [51] mention some potential reasons and measures to improve this situation, namely: difficulty on performing experiments where humans are involved, the lack of common benchmarks, or even that empirical work is not encouraged by the journals and conferences of this area.

3) RQ3. HOW FAR ARE WE FROM A TOOL THAT SOLVES OUR THREE REQUIREMENTS?

In figure 5 we display a Venn diagram with our three requirements. We can see there is a single study that covers the three requirements: *A Distributed In-situ Analysis Method*

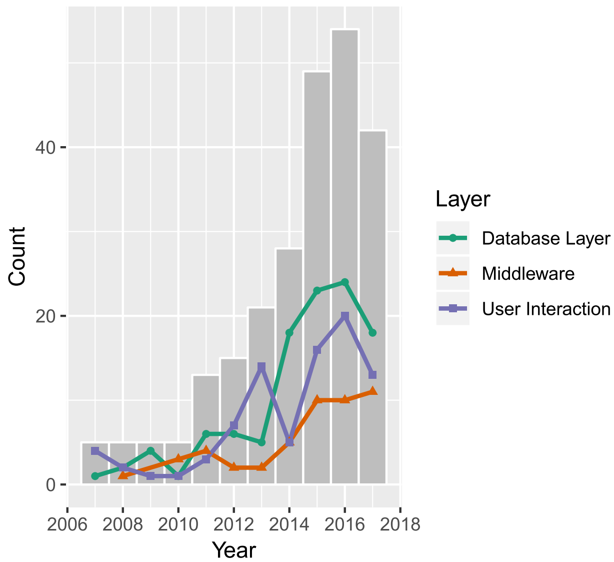

FIGURE 4. Number of papers per layer and year. Note that the drop during 2017 is due to the search having been done in May 2017.

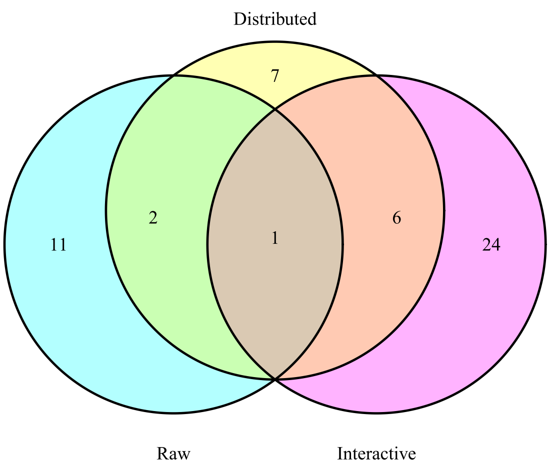

FIGURE 5. Venn diagram with solutions that satisfy our requirements.

for Large-scale Scientific Data, by Han *et al.* [54]. While they mention the access over raw files and the fact that it is distributed, they do not explicitly state anything about their interactivity. However, the measured times for selective queries that they report are in the order of a few seconds. Consequently, we decided to consider it to be suitable for interactive usage.

The tests they perform use datasets that are close to the memory available on the system and, therefore, more tests with bigger dataset sizes could be needed.

Aside from this paper, no other study combines access to raw data with low response times.

The solutions that cover at least two out of the three requirements are summarized in more detail in section V.

B. STUDY INSIGHTS

Research in data exploration is very active and there has been—and there is—a myriad of solutions proposed. In fact, this should not come as a surprise: in 2005 Stonebraker and Cetintemel [55] had already predicted this was bound to happen and predicted that there would be an increase of domain-specific tools. This would explain why, of the all classified studies, only one tool satisfies our three prerequisites.

In general, several different systems and approaches have been proposed, which could, perhaps, be seen as building blocks. Not all combinations necessarily make sense but it seems that there are research opportunities in this direction, depending on the specific needs to be covered.

For instance, in our particular case, we could consider combining distributed access over raw files, as [54] does, but using approximate query processing to reduce the response times.

Code generation is a popular approach for querying raw data files and approximation-aware code generation has been noted as a challenge that is yet to be addressed [31]. Consequently, more work on this particular overlap of approaches may provide interesting results.

On a orthogonal consideration, since the generation of data volume will likely not slow down, the trend for more tools covering specific niches is probably going to continue. This diversity of tools is a challenge in itself in many respects, for example: How do we choose the right solution? What is the cost of making the wrong choice? What happens if the chosen tool goes unmaintained in the future and there is no community around it? Will it be hard to maintain? Of course, these questions are not new in software engineering but typically there are not many choices when it comes to decide on traditional data storage systems, such as DBMS. In the last decade, there has been an increase of available options (relational, object oriented, schema-less, key-value, ...) and, while opting for a DBMS has become harder, it has remained rather manageable. However, looking at the results of this study, the difficulty for users to decide will likely become more challenging.

C. THREATS TO VALIDITY

1) SEARCH BIAS

The gaps identified may be covered in journals and conferences associated with the user domain—e.g. astrophysics—, rather than with computer science and engineering. The forward snowballing step reduces this risk because these hypothetical publications would most likely cite the original proposal of solution. However, considering that our research method has allowed us to find even gray literature, we consider this risk to be low.

2) FILTERING OF ARTICLES

Given the huge number of papers that resulted from the search, a first filtering was done just based on title and abstract. This is a difficult challenge. Unlike in other disciplines, sometimes abstracts do not contain enough information about the paper and keywords can be inconsistent between journals and authors [40], [45], [56]. As recommended by [45], we have also taken into consideration the conclusions to cover this issue.

3) CLASSIFICATION

Another concern about these classifications is the bias of the researcher's own interpretation [57]. For instance, Jorgensen and Shepperd [43] report on a disagreement over 39% of the reviewed papers in their systematic review due to different interpretations of the description of each category. We have been careful in this respect to guarantee the internal validity of the study, although some misclassification may still exist.

Additionally, it can be hard to identify if a solution covers or not one of the three predefined requirements based just on a paper. They may not have been explicitly mentioned if the authors did not consider them relevant for the purposes of their publication. Therefore, there may have been false negatives.

The present paper documents our process and the resulting publication list has been made publicly available—see subsection II-C—, so anyone interested can replicate and/or validate our results.

V. DISCUSSION OF RELEVANT METHODS

Included for completeness is a summary of each of the nine publications that cover, at least, two out of the three requirements.

A. ALL THREE REQUIREMENTS

As already mentioned, the only solution that covers the three requirements is documented on the paper "A Distributed In-situ Analysis Method for Large-scale Scientific Data" [54], classified as "adaptive loading".

Stonebraker *et al.* [58] build on top of SciDB, a distributed array-based scientific database, and focus on HDF files [59]. To avoid the overhead of data pre-loading, they leverage the flexible architecture of this database engine, providing their own scan operator to read the data directly from the raw files when needed, which needs to be adapted to the internal representation of SciDB.

This adaptation is done in two different stages: local and global mapping.

During the local mapping, they read on demand the data that matches the filters associated to the query, adapting it to the SciDB chunk representation: pieces of array data that are distributed together based on some policy - e.g hashing, range partitioning.

At the global mapping stage, the resulting chunks are redistributed across the storage nodes following the SciDB policies.

Although not relevant for our use case, it is worth mentioning that they also merge small files together to reduce the performance penalty of processing many small files.

This approach is interesting as it compartmentalizes well the logic required to access the raw data from the file distribution and the query engine.

However, the paper notably misses information about the network traffic caused by their global mapping stage, since the network overhead depends on how the actual data distribution matches SciDB expectations.

B. DISTRIBUTED ACCESS TO RAW FILES

DiNoDB [60] is oriented towards the interactive development of data aggregation algorithms, where the user needs to move quickly between the batch processing stage and the interactive evaluation of the quality of the results.

It is deployed together with Hadoop and it generates the auxiliary metadata using user defined functions executed by the reducers during the batch processing stage. Therefore, the metadata ends up stored together with the raw data - the output of the reducers, and will also be replicated by the Hadoop Distributed File System (HDFS) across the cluster. Additionally, the output data may be cached optionally in memory - via ramfs or the filesystem cache.

For the interactive stage, on each HDFS Data Node it is deployed an instance of a customized PostgresRaw [17] database, a modified version of PostgreSQL with additional support for raw files based on positional maps - positions of attributes within the file.

With this architecture deployment, the client 1) issues the query to each node separately; 2) PostgresRaw uses the indices to retrieve the offsets of the relevant records and the positional maps to find the fields within the raw file; and 3) the client aggregates the results.

This approach gets good response times for the interactive stage, but the latency increases significantly when the output data does not fully fit into memory.

ARMFUL (Analysis of Raw data from Multiple Files) [61], probably has the most strict requirement set of all the analyzed papers. Its authors need to access raw data generated during the execution of a workflow and collect their provenance with high granularity. While other tools keep track of the data provenance at the file level - leaving to the user the cross-match of records stored in different files - they are able to associate related data entries contained in the raw data files at the record level.

To do so, the authors formally define two additional workflow algebraic data operators [62], which allows to address specific records stored on a file within a dataflow: *Raw Data Extraction* - read, tokenize, filter, parse - and *Raw Data Indexing*. These operators can be composed with the existing ones, as *Map* or *Filter* - for instance, a user could map a list of file names to their content and then filter records with a specific threshold, keeping track of the provenance of the data during all the process.

The indexing can rely on external tools, and two implementations are provided: one based on bitmap indexes generated by FastBit [63], and another one on positional maps, implemented following RAW's approach [9].

Since this study focus particularly on raw data access during simulations, the interactivity only applies to the queries made to the provenance database.

C. DISTRIBUTED AND INTERACTIVE

This combination is the one with the most matching methods. Five out of the six ones are classified as "query approximation", and the remaining one, even though labeled as "visual optimization", relies heavily on query approximation as well.

It would seem that to get fast responses some compromises on the precision have to be made. This makes sense intuitively as processing less data will reduce the processing time at the cost of less accuracy. Additionally, on a distributed system, some nodes may be offline, unresponsive or overloaded. In order to keep the latency low, the results need to be aggregated within a reasonable deadline, even if parts of the system have not responded yet.

It is worth noting that most of these papers also match the "sampling" category, but since sampling is just an aspect of the overall solution and their authors normally use "query approximation" to refer to their methods, we have decided to classify them as such.

BlinkDB [64] allows users to perform SQL-like aggregation queries on data stored on HDFS, specifying time or error constraints. First, the authors base their system on the assumption - supported by evidence - that the column sets used for the aggregation queries are predictable, regardless of the actual grouping value. With this information, they perform a stratified sampling [65] to avoid the under-representation of rare subgroups. Finally, the system chooses the suitable samples based on the query constraints provided by the user, profiling them at run time so it can improve the execution plan for later queries.

ScalaR [66] improves the performance of the visualization of big data sets dynamically reducing the size of the response returned to the front-end layer. Its authors provide an intermediate layer that consumes the queries issued by the user and uses the statistics computed by the database backend to evaluate in advance the expected size of the result set. If this size is above a given threshold, the query is rewritten to either aggregate, sample or filter the data, generating a smaller approximate response that can be displayed more performantly.

Although their solution is back-end agnostic, their proposed implementation relies on SciDB [58]. It quickly comes to mind that this could potentially be integrated with the previous method by Han *et al*. [54], resulting on a visual exploration tool for raw data files.

The authors of **DICE** (Distributed and Interactive Cube Exploration) [67] attack the problem on three fronts: speculative query execution, online data sampling, and an exploration model - *faceted* cube exploration - that limits the number of possible queries, improving the efficacy of the speculative execution.

Probably, the most interesting idea from this paper is the notion of the exploration being done in "sessions": The authors do not attempt to optimize for any possible query, but only for those that are likely to follow from the state of the current session. Predicting a set of potential following queries is made possible thanks to their exploration model, which restricts the possible number of "transitions" from the current state for a session.

The predicted queries are then ranked based on their likelihood and accuracy gain, and those that are most likely and provide the most accuracy gain will be speculatively executed in advance, populating the cache. When the final query arrives, the response can be built from the content of the cache if the predictions were successful. Otherwise, it will be scheduled to the underlying nodes.

For more information about "data cubes", we refer to the DICE paper, or the original proposal [68].

AccuracyTrader [69] is a distributed approximate processing system comprised of two components: one online and one offline.

First, the offline part reduces the dimensionality of the original data using Single Value Decomposition - so it only supports numerical values. Then, it groups similar entries using an R-Tree, where each node represents an aggregated data point, and all nodes at the same level correspond to a "synopsis". This tree is flattened into an index at a level that balances between the number of leaves under each aggregated data point and the selectivity of the tree at that level. Finally, it aggregates the data for each index entry using the original dimensions of the indexed points and stores this aggregated data into the "synopsis".

When a query arrives, the online part uses these "synopsis" to produce an approximate result with an accuracy estimation. It then iterates using the detailed data points to improve the response accuracy until the deadline specified by the user expires.

In this paper, the authors prove that the system scales well in terms of tail latency and accuracy when the number of requests increases for a "search engine"-like workload. However, the data has to be aggregated into the synopsis beforehand.

KIWI [70] is a SQL front-end built on top of Hadoop that aims to provide both batch processing and interactive analytics via approximate query processing. It generates both vertical (column) and horizontal (row) samples, and re-writes the queries to use these samples instead of the original data. However, it is hard to assess the technical soundness of this solution, since the paper is very short - 2 pages including citations - and we have not been able to find any later citations nor do the authors cite other papers about the same tool.

Finally, Wang *et al*. [71] introduce a framework based on the map-reduce paradigm. Instead of the traditional batch processing approach where the analysis is performed on big chunks of data, their system executes the analysis logic iteratively on samples, updating an estimator in each round until a stop condition is satisfied - both estimator and condition provided by the user. When the termination condition is satisfied, the remaining jobs are canceled, saving computing cycles and reducing the latency. Similarly to other analyzed solutions, they use a stratified sampling to ensure a good accuracy and the coverage of rare cases. The sampling is done

Title	Year	Cluster	Type	Ref.
A Discussion on Visual Interactive Data Exploration Using Self-Organizing Maps	2011	Visualization Tools	Validation Research	[72]
A Distributed Infrastructure for Earth-Science Big Data Retrieval	2015	Novel Query Interfaces	Proposal of Solution	[73]
A GPU-based index to support interactive spatio-temporal queries over historical data	2016	Spatial Query	Proposal of Solution	[74]
A Holistic Approach to OLAP Sessions Composition	2014	Assisted Query Formulation	Proposal of Solution	[75]
A Logic-Based Approach to Mining Inductive Databases	2007	Novel Query Interfaces	Proposal of Solution	[76]
A Scalable Execution Engine for Package Queries	2017	Novel Query Interfaces	Proposal of Solution	[77]
A Schema-Based Approach to Enable Data Integration on the Fly	2017	Flexible Engines	Proposal of Solution	[78]
A Signaling Game Approach to Databases Querying and Interaction	2016	Novel Query Interfaces	Proposal of Solution	[79]
A Unified Correlation-based Approach to Sampling Over Joins	2017	Sampling	Proposal of Solution	[80]
A distributed in-situ analysis method for large-scale scientific data	2017	Adaptive Loading	Proposal of Solution	[81]
A framework for query refinement with user feedback	2013	Assisted Query Formulation	Proposal of Solution	[82]
A graphical system for interactive creation and exploration of dynamic information visualization	2016	Visualization Tools	Proposal of Solution	[83]
A hierarchical aggregation framework for efficient multilevel visual exploration and analysis	2017	Adaptive Indexing	Proposal of Solution	[84]
A study of SQL-on-Hadoop systems	2014	Exploration Interfaces	Validation Research	[85]
A taxonomy for region queries in spatial databases	2015	Spatial Query	Evaluation Research	[86]
A time-series compression technique and its application to the smart grid	2015	Time Series	Proposal of Solution	[87]
ADS: the adaptive data series index	2016	Adaptive Indexing	Proposal of Solution	[88]
AIDE: An Active Learning-Based Approach for Interactive Data Exploration	2016	Sampling	Proposal of Solution	[89]
AIR: Adaptive Index Replacement in Hadoop	2015	Adaptive Indexing	Proposal of Solution	[90]
AQP++: A Hybrid Approximate Query Processing Framework for Generalized Aggregation Queries	2017	Query Approximation	Proposal of Solution	[91]
AQUAdexIM: highly efficient in-memory indexing and querying of astronomy time series images	2016	Time Series	Proposal of Solution	[92]
About Database Summarization	2010	Query Approximation	Proposal of Solution	[93]
Abstraction Without Regret in Database Systems Building: a Manifesto	2014	Flexible Engines	Philosophical Paper	[94]
Access Path Selection in Main-Memory Optimized Data Systems: Should I Scan or Should I Probe?	2017	Indexes	Evaluation Research	[95]
AccuracyTrader: Accuracy-Aware Approximate Processing for Low Tail Latency and High Result Accuracy in Cloud Online Services	2016	Query Approximation	Proposal of Solution	[96]
Adaptive Indexing over Encrypted Numeric Data	2016	Adaptive Indexing	Proposal of Solution	[97]
Adaptive indexing approach for main memory column store	2016	Adaptive Indexing	Proposal of Solution	[98]
Adaptive query processing on RAW data	2014	Flexible Engines	Proposal of Solution	[99]
Adaptive-sampling algorithms for answering aggregation queries on Web sites	2008	Sampling	Validation Research	[100]
Alpine: Efficient In-Situ Data Exploration in the Presence of Updates	2017	Adaptive Indexing	Proposal of Solution	[101]
An Adaptive Data Partitioning Scheme for Accelerating Exploratory Spark SQL Queries	2017	Adaptive Storage	Proposal of Solution	[102]

An Analysis of Query-Agnostic Sampling for Interactive Data Exploration	2017	Automatic Exploration	Evaluation Research	[103]
An Efficient Time Optimized Scheme for Progressive Analytics in Big Data	2015	Query Approximation	Proposal of Solution	[104]
An enhanced visualization process model for incremental visualization	2016	Visual Optimizations	Proposal of Solution	[105]
An experimental evaluation and analysis of database cracking	2016	Adaptive Indexing	Evaluation Research	[106]
An intelligent, uncertainty driven aggregation scheme for streams of ordered sets	2016	Query Approximation	Proposal of Solution	[107]
Analytics in Motion: High Performance Event-Processing AND Real-Time Analytics in the Same Database	2015	Adaptive Storage	Proposal of Solution	[108]
Answering Temporal Analytic Queries over Big Data Based on Precomputing Architecture	2017	Time Series	Proposal of Solution	[109]
Approximate OLAP on Sustained Data Streams	2017	Query Approximation	Proposal of Solution	[110]
Approximate Query Engines : Commercial Challenges and Research Opportunities	2017	Query Approximation	Opinion Papers	[111]
Approximate Query Processing: No Silver Bullet	2017	Query Approximation	Evaluation Research	[112]
Approximate range searching in external memory	2011	Query Approximation	Proposal of Solution	[113]
AstroShelf: understanding the universe through scalable navigation of a galaxy of annotations	2012	Visualization Tools	Proposal of Solution	[114]
Benchmarking exploratory OLAP	2017	Assisted Query Formulation	Validation Research	[115]
Beyond one billion time series: Indexing and mining very large time series collections with iSAX2+	2014	Time Series	Evaluation Research	[116]
Beyond the Wall: Near-Data Processing for Databases	2015	Adaptive Loading	Proposal of Solution	[117]
Bi-Level Online Aggregation on Raw Data	2017	Sampling	Proposal of Solution	[118]
Big sequence management: A glimpse of the past, the present, and the future	2016	Time Series	Validation Research	[119]
BlinkDB: queries with bounded errors and bounded response times on very large data	2013	Query Approximation	Proposal of Solution	[120]
Bridging the Archipelago between Row-Stores and Column-Stores for Hybrid Workloads	2016	Flexible Engines	Proposal of Solution	[121]
Building efficient query engines in a high-level language	2014	Flexible Engines	Proposal of Solution	[122]
Cell-at-a-Time Approach to Lazy Evaluation of Dimensional Aggregations	2013	Query Approximation	Proposal of Solution	[123]
Cheetah: a high performance, custom data warehouse on top of MapReduce	2010	Data Prefetching	Proposal of Solution	[124]
CliffGuard: A Principled Framework for Finding Robust Database Designs	2015	Flexible Engines	Proposal of Solution	[125]
Cluster-Driven Navigation of the Query Space	2016	Novel Query Interfaces	Proposal of Solution	[126]
Clustrophile: A Tool for Visual Clustering Analysis	2016	Visualization Tools	Proposal of Solution	[127]
Combining Design and Performance in a Data Visualization Management System	2017	Visualization Tools	Proposal of Solution	[128]
Computer-Assisted Query Formulation	2016	Assisted Query Formulation	Evaluation Research	[129]
Concurrency control for adaptive indexing	2012	Adaptive Indexing	Proposal of Solution	[130]
Controlling False Discoveries During Interactive Data Exploration	2017	Visual Optimizations	Proposal of Solution	[131]
D-Ocean: an unstructured data management system for data ocean environment	2016	Flexible Engines	Proposal of Solution	[132]
DAQ: A New Paradigm for Approximate Query Processing	2015	Query Approximation	Proposal of Solution	[133]
DBMS Data Loading: An Analysis on Modern Hardware	2017	Adaptive Loading	Evaluation Research	[134]

DIRAQ: scalable in situ data- and resource-aware indexing for optimized query performance	2014	Adaptive Indexing	Proposal of Solution	[135]
Data Canopy: Accelerating Exploratory Statistical Analysis	2017	Data Prefetching	Proposal of Solution	[136]
Data Exploration with SQL using Machine Learning Techniques	2017	Assisted Query Formulation	Proposal of Solution	[137]
Data Tweening: Incremental Visualization of Data Transforms	2017	Visual Optimizations	Proposal of Solution	[138]
Data series management: The road to big sequence analytics	2015	Indexes	Evaluation Research	[139]
Data vaults: A symbiosis between database technology and scientific file repositories	2012	Adaptive Loading	Proposal of Solution	[140]
DataPlay: interactive tweaking and example-driven correction of graphical database queries	2012	Assisted Query Formulation	Proposal of Solution	[141]
Database Cracking: Fancy Scan, Not Poor Man's Sort!	2014	Adaptive Indexing	Proposal of Solution	[142]
Database Learning: Toward a Database that Becomes Smarter Every Time	2017	Query Approximation	Philosophical Paper	[143]
Delay aware querying with Seaweed	2008	Query Approximation	Proposal of Solution	[144]
Deterministic View Selection for Data-Analysis Queries: Properties and Algorithms	2012	Data Prefetching	Evaluation Research	[145]
DiNoDB: Efficient Large-Scale Raw Data Analytics	2014	Adaptive Indexing	Proposal of Solution	[146]
Discovering Queries Based on Example Tuples	2014	Assisted Query Formulation	Proposal of Solution	[147]
Distributed and interactive cube exploration	2014	Sampling	Proposal of Solution	[148]
DivIDE: Efficient Diversification for Interactive Data Exploration	2014	Data Prefetching	Proposal of Solution	[149]
Diversifying with Few Regrets, But too Few to Mention	2015	Query Approximation	Proposal of Solution	[150]
Does Online Evaluation Correspond to Offline Evaluation in Query Auto Completion?	2017	Assisted Query Formulation	Evaluation Research	[151]
Dynamic Prefetching of Data Tiles for Interactive Visualization	2016	Data Prefetching	Proposal of Solution	[152]
Dynamic reduction of query result sets for interactive visualizaton	2013	Visual Optimizations	Proposal of Solution	[153]
Efficient Evaluation of Object-Centric Exploration Queries for Visualization	2015	Visual Optimizations	Proposal of Solution	[154]
Efficient schemes for similarity-aware refinement of aggregation queries	2017	Query Approximation	Proposal of Solution	[155]
End-User Development of Information Visualization	2013	Visual Optimizations	Evaluation Research	[156]
Enhanced Query-by-Object approach for information requirement elicitation in large databases	2012	Novel Query Interfaces	Proposal of Solution	[157]
Enhancing Parallel Data Loading for Large Scale Scientific Database	2015	Adaptive Loading	Proposal of Solution	[158]
Evaluating a Stream of Relational K NN Queries by a Knowledge Base	2015	Data Prefetching	Proposal of Solution	[159]
Exact indexing for massive time series databases under time warping distance	2010	Time Series	Proposal of Solution	[160]
Exemplar queries: a new way of searching	2016	Novel Query Interfaces	Proposal of Solution	[161]
Exploring Databases via Reverse Engineering Ranking Queries with PALEO	2016	Automatic Exploration	Proposal of Solution	[162]
Fast and adaptive indexing of multi-dimensional observational data	2016	Adaptive Indexing	Proposal of Solution	[163]
Fast queries over heterogeneous data through engine customization	2016	Flexible Engines	Proposal of Solution	[164]
Fast, Explainable View Detection to Characterize Exploration Queries	2016	Assisted Query Formulation	Proposal of Solution	[165]
Fast-Forwarding to Desired Visualizations with zenvisage	2017	Visualization Tools	Proposal of Solution	[166]

FlashExtract: a framework for data extraction by examples	2013	Assisted Query Formulation	Validation Research	[167]
Flying KIWI: Design of Approximate Query Processing Engine for Interactive Data Analytics at Scale	2015	Query Approximation	Proposal of Solution	[168]
Gestural query specification	2013	Novel Query Interfaces	Philosophical Paper	[169]
H2O: A Hands-free Adaptive Store	2014	Adaptive Storage	Proposal of Solution	[170]
Hashedcubes: Simple, Low Memory, Real-Time Visual Exploration of Big Data	2017	Spatial Query	Proposal of Solution	[171]
Holistic Indexing in Main-memory Column-stores	2015	Adaptive Indexing	Proposal of Solution	[172]
How Progressive Visualizations Affect Exploratory Analysis	2017	Query Approximation	Validation Research	[173]
IEVQ: An Iterative Example-Based Visual Query for Pathology Database	2017	Novel Query Interfaces	Proposal of Solution	[174]
IVIS4BigData: A reference model for advanced visual interfaces supporting big data analysis in virtual research environments	2016	Visual Optimizations	Philosophical Paper	[175]
IncApprox: A Data Analytics System for Incremental Approximate Computing	2016	Query Approximation	Proposal of Solution	[176]
Indexing for interactive exploration of big data series	2014	Time Series	Proposal of Solution	[177]
Information retrieval using dynamic indexing	2014	Adaptive Indexing	Proposal of Solution	[178]
Initial Sampling for Automatic Interactive Data Exploration	2016	Sampling	Proposal of Solution	[179]
Intelligent Data Granulation on Load: Improving Infobright's Knowledge Grid	2009	Adaptive Storage	Proposal of Solution	[180]
Interactive Browsing and Navigation in Relational Databases	2016	Visualization Tools	Proposal of Solution	[181]
Interactive Data Exploration Using Semantic Windows	2014	Data Prefetching	Proposal of Solution	[182]
Interactive Inference of Join Queries	2014	Assisted Query Formulation	Evaluation Research	[183]
Interactive SQL query suggestion: Making databases user-friendly	2011	Assisted Query Formulation	Proposal of Solution	[184]
Interactive Visualization of Big Data	2016	Visual Optimizations	Proposal of Solution	[185]
Interactive and Scalable Exploration of Big Spatial Data – A Data Management Perspective	2015	Spatial Query	Philosophical Paper	[186]
Interactive time series exploration powered by the marriage of similarity distances	2016	Time Series	Proposal of Solution	[187]
Invisible Glue : Scalable Self-Tuning Multi-Stores	2015	Flexible Engines	Proposal of Solution	[188]
Invisible loading	2013	Adaptive Loading	Proposal of Solution	[189]
Keyword Search in Relational Databases: A Survey	2010	Assisted Query Formulation	Evaluation Research	[190]
Knowing When You're Wrong: Building Fast and Reliable Approximate Query Processing Systems	2014	Query Approximation	Proposal of Solution	[191]
Kodiak: leveraging materialized views for very low-latency analytics over high-dimensional web-scale data	2016	Data Prefetching	Evaluation Research	[192]
L-Store: A Real-time OLTP and OLAP System	2016	Adaptive Storage	Proposal of Solution	[193]
Learning Path Queries on Graph Databases	2015	Automatic Exploration	Proposal of Solution	[194]
Learning Queries from Examples and Their Explanations	2016	Automatic Exploration	Proposal of Solution	[195]
Learning and verifying quantified boolean queries by example	2013	Assisted Query Formulation	Proposal of Solution	[196]
Logic-Partition Based Gaussian Sampling for Online Aggregation	2017	Sampling	Proposal of Solution	[197]
Main Memory Adaptive Indexing for Multi-core Systems	2014	Adaptive Indexing	Proposal of Solution	[198]
Managing Massive Time Series Streams with Multi-Scale Compressed Trickles	2009	Time Series	Proposal of Solution	[199]
Meet Charles, big data query advisor	2013	Assisted Query Formulation	Proposal of Solution	[200]
Merging file systems and data bases to fit the grid	2010	Data Prefetching	Proposal of Solution	[201]

Merging what's cracked, cracking what's merged	2011	Adaptive Indexing	Proposal of Solution	[202]
Merlin: Exploratory Analysis with Imprecise Queries	2016	Assisted Query Formulation	Proposal of Solution	[203]
Model-Based Diversification for Sequential Exploratory Queries	2017	Data Prefetching	Proposal of Solution	[204]
Model-driven Visual Analytics	2008	Visual Optimizations	Validation Research	[205]
Modeling Large Time Series for Efficient Approximate Query Processing	2015	Query Approximation	Proposal of Solution	[206]
Modeling Semantic and Behavioral Relations for Query Suggestion	2013	Assisted Query Formulation	Proposal of Solution	[207]
MuVE: Efficient Multi-Objective View Recommendation for Visual Data Exploration	2016	Visualization Tools	Proposal of Solution	[208]
NoDB: efficient query execution on raw data files	2012	Adaptive Indexing	Proposal of Solution	[209]
ORange: Objective-Aware Range Query Refinement	2014	Query Approximation	Proposal of Solution	[210]
On Improving User Response Times in Tableau	2015	Visual Optimizations	Proposal of Solution	[211]
On Interactive Pattern Mining from Relational Databases	2007	Assisted Query Formulation	Proposal of Solution	[212]
On query result diversification	2011	Data Prefetching	Proposal of Solution	[213]
On the analysis of big data indexing execution strategies	2017	Indexes	Evaluation Research	[214]
Optimized Disk Layouts for Adaptive Storage of Interaction Graphs	2014	Adaptive Storage	Proposal of Solution	[215]
Optimized Multi-Resolution Indexing and Retrieval Scheme of Time Series	2015	Time Series	Validation Research	[216]
Optimizing database load and extract for big data era	2014	Adaptive Loading	Proposal of Solution	[217]
Organic databases	2011	Flexible Engines	Evaluation Research	[218]
PABIRS: A data access middleware for distributed file systems	2015	Adaptive Indexing	Proposal of Solution	[219]
PFunk-H: approximate query processing using perceptual models.	2016	Query Approximation	Proposal of Solution	[220]
Past and Future Steps for Adaptive Storage Data Systems: From Shallow to Deep Adaptivity	2016	Adaptive Storage	Opinion Papers	[221]
Progressive diversification for column-based data exploration platforms	2015	Adaptive Loading	Proposal of Solution	[222]
QPlain: Query by explanation	2016	Automatic Exploration	Proposal of Solution	[223]
QueRIE reloaded: Using matrix factorization to improve database query recommendations	2015	Assisted Query Formulation	Proposal of Solution	[224]
Query Similarity for Approximate Query Answering	2016	Query Approximation	Validation Research	[225]
Query Workloads for Data Series Indexes	2015	Indexes	Evaluation Research	[226]
Query by output	2009	Assisted Query Formulation	Proposal of Solution	[227]
Query from examples: An iterative, data-driven approach to query construction	2015	Automatic Exploration	Proposal of Solution	[228]
Querying Big Data by Accessing Small Data	2015	Query Approximation	Proposal of Solution	[229]
Querying Time Interval Data	2015	Time Series	Proposal of Solution	[230]
Querying continuous functions in a database system	2008	Novel Query Interfaces	Proposal of Solution	[231]
Quickr: Lazily Approximating Complex AdHoc Queries in BigData Clusters	2016	Query Approximation	Proposal of Solution	[232]
R-proxy framework for in-DB data-parallel analytics	2012	Novel Query Interfaces	Proposal of Solution	[233]
RDQS: A Relevant and Diverse Query Suggestion Generation Framework	2015	Assisted Query Formulation	Proposal of Solution	[234]
REQUEST: A scalable framework for interactive construction of exploratory queries	2016	Assisted Query Formulation	Proposal of Solution	[235]
RailwayDB: adaptive storage of interaction graphs	2016	Adaptive Storage	Proposal of Solution	[236]
Rapid Sampling for Visualizations with Ordering Guarantees	2015	Visual Optimizations	Proposal of Solution	[237]
Raw data queries during data-intensive parallel workflow execution	2016	Adaptive Indexing	Proposal of Solution	[238]

Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning	2016	Adaptive Indexing	Proposal of Solution	[239]
Research and application of query rewriting based on materialized views	2011	Data Prefetching	Evaluation Research	[240]
Resilient store: A heuristic-based data format selector for intermediate results	2016	Adaptive Storage	Proposal of Solution	[241]
Revisiting Reuse for Approximate Query Processing	2017	Query Approximation	Proposal of Solution	[242]
S4: Top-k Spreadsheet-Style Search for Query Discovery	2015	Assisted Query Formulation	Proposal of Solution	[243]
SCANRAW: A Database Meta-Operator for Parallel In-Situ Processing and Loading	2015	Adaptive Loading	Proposal of Solution	[244]
SCOUT: Prefetching for Latent Structure Following Queries	2012	Data Prefetching	Proposal of Solution	[245]
SOCR data dashboard: an integrated big data archive mashing medicare, labor, census and econometric information	2015	Visualization Tools	Validation Research	[246]
STORM : Spatio-Temporal Online Reasoning and Management of Large Spatio-Temporal Data	2015	Query Approximation	Proposal of Solution	[247]
Sample + Seek : Approximating Aggregates with Distribution Precision Guarantee	2016	Query Approximation	Proposal of Solution	[248]
Sampling for scalable visual analytics	2017	Sampling	Evaluation Research	[249]
Scaling Up Mixed Workloads: A Battle of Data Freshness, Flexibility, and Scheduling	2015	Flexible Engines	Evaluation Research	[250]
Scaling and time warping in time series querying	2008	Time Series	Proposal of Solution	[251]
SciBORQ: Scientific data management with Bounds On Runtime and Quality	2011	Sampling	Proposal of Solution	[252]
Scientific discovery through weighted sampling	2013	Sampling	Proposal of Solution	[253]
Searchlight: Enabling Integrated Search and Exploration over Large Multidimensional Data	2015	Flexible Engines	Proposal of Solution	[254]
SeeDB: Visualizing Database Queries Efficiently	2013	Visual Optimizations	Proposal of Solution	[255]
Self-Driving Database Management Systems	2017	Adaptive Storage	Proposal of Solution	[256]
Self-organizing Tuple Reconstruction in Column-stores	2009	Adaptive Indexing	Proposal of Solution	[257]
Semi-Automated Exploration of Data Warehouses	2015	Assisted Query Formulation	Proposal of Solution	[258]
Skipping-oriented Partitioning for Columnar Layouts	2016	Adaptive Storage	Proposal of Solution	[259]
Slalom : Coasting Through Raw Data via Adaptive Partitioning and Indexing	2017	Adaptive Indexing	Proposal of Solution	[260]
SnapToQuery: Providing Interactive Feedback during Exploratory Query Specification	2015	Assisted Query Formulation	Proposal of Solution	[261]
Spatial online sampling and aggregation	2015	Sampling	Proposal of Solution	[262]
Speed Up Distance-Based Similarity Query Using Multiple Threads	2014	Spatial Query	Validation Research	[263]
Stale View Cleaning: Getting Fresh Answers from Stale Materialized Views	2015	Sampling	Proposal of Solution	[264]
Stochastic Database Cracking: Towards Robust Adaptive Indexing in Main-Memory Column-Stores	2012	Adaptive Indexing	Proposal of Solution	[265]
Supporting online analytics with user-defined estimation and early termination in a MapReduce-like framework	2015	Query Approximation	Proposal of Solution	[266]
Symbolic representation of time series: A hierarchical coclustering formalization	2016	Time Series	Proposal of Solution	[267]
Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches	2011	Query Approximation	Proposal of Solution	[268]
The Analytical Bootstrap: A New Method for Fast Error Estimation in Approximate Query Processing	2014	Query Approximation	Proposal of Solution	[269]
The Case for Data Visualization Management Systems [Vision Paper]	2012	Visual Optimizations	Proposal of Solution	[270]

The Case for RodentStore, an Adaptive, Declarative Storage System	2009	Flexible Engines	Proposal of Solution	[271]
The Researcher's Guide to the Data Deluge: Querying a Scientific Database in Just a Few Seconds	2011		Philosophical Paper	[272]
The array database that is not a database: File based array query answering in rasdaman	2013	Adaptive Loading	Proposal of Solution	[273]
The case for multi-engine data analytics	2014	Flexible Engines	Philosophical Paper	[274]
The design of an adaptive column-store system	2017	Adaptive Storage	Evaluation Research	[275]
The power of choice in data-aware cluster scheduling	2014	Sampling	Proposal of Solution	[276]
The uncracked pieces in database cracking	2013	Adaptive Indexing	Validation Research	[277]
Time series subsequence matching based on a combination of PIP and clipping	2011	Time Series	Proposal of Solution	[278]
TimeExplorer: Similarity search time series by their signatures	2013	Time Series	Proposal of Solution	[279]
TimeLine and visualization of multiple-data sets and the visualization querying challenge	2007	Visual Optimizations	Proposal of Solution	[280]
Towards Best Region Search for Data Exploration	2016	Spatial Query	Proposal of Solution	[281]
Towards a One Size Fits All Database Architecture	2011	Adaptive Storage	Proposal of Solution	[282]
Towards a scalable, performance-oriented OLAP storage engine	2012	Flexible Engines	Proposal of Solution	[283]
Towards an Adaptive Framework for Real-Time Visualization of Streaming Big Data	2017	Visual Optimizations	Proposal of Solution	[284]
Towards an efficient storage and retrieval mechanism for large unstructured grids	2015	Data Prefetching	Proposal of Solution	[285]
Towards zero-overhead static and adaptive indexing in Hadoop	2014	Adaptive Indexing	Proposal of Solution	[286]
Trust, but Verify : Optimistic Visualizations of Approximate Queries for Exploring Big Data	2017	Visual Optimizations	Proposal of Solution	[287]
Tsdb: A compressed database for time series	2012	Time Series	Proposal of Solution	[288]
Updating a cracked database	2007	Adaptive Indexing	Proposal of Solution	[289]
User Interaction Models for Disambiguation in Programming by Example	2015	Assisted Query Formulation	Proposal of Solution	[290]
User search intention in interactive data exploration: A brief review	2017	Assisted Query Formulation	Evaluation Research	[291]
User's interpretations of features in visualization	2015	Visual Optimizations	Proposal of Solution	[292]
User-driven refinement of imprecise queries	2014	Assisted Query Formulation	Proposal of Solution	[293]
Using Information Visualization to support Open Data Integration	2015	Visual Optimizations	Evaluation Research	[294]
VDDA: automatic visualization-driven data aggregation in relational databases	2016	Visual Optimizations	Proposal of Solution	[295]
Vertical partitioning for query processing over raw data	2015	Flexible Engines	Proposal of Solution	[296]
Visual Analytics in Environmental Research: A Survey on Challenges, Methods and Available Tools	2013	Visual Optimizations	Philosophical Paper	[297]
Visual Data Exploration Using Webbles	2013	Visual Optimizations	Proposal of Solution	[298]
Visual exploration of machine learning results using data cube analysis	2016	Visualization Tools	Proposal of Solution	[299]
Visual query specification and interaction with industrial engineering data	2013	Novel Query Interfaces	Proposal of Solution	[300]
Visual reasoning indexing and retrieval using in-memory computing	2016	Visual Optimizations	Evaluation Research	[301]
Visualization-aware sampling for very large databases	2016	Sampling	Proposal of Solution	[302]
Visualizing Big Data with augmented and virtual reality: challenges and research agenda	2015	Visual Optimizations	Evaluation Research	[303]
Visually defining and querying consistent multi-granular clinical temporal abstractions	2012	Visual Optimizations	Proposal of Solution	[304]

VizDeck: self-organizing dashboards for visual analytics	2012	Visualization Tools	Proposal of Solution	[305]
What Users Don't Expect about Exploratory Data Analysis on Approximate Query Processing Systems	2017	Query Approximation	Validation Research	[306]
Wide Table Layout Optimization based on Column Ordering and Duplication	2017	Adaptive Storage	Proposal of Solution	[307]
Workload-Driven Antijoin Cardinality Estimation	2015	Sampling	Proposal of Solution	[308]
XmdvtoolQ:: Quality-aware Interactive Data Exploration	2007	Novel Query Interfaces	Proposal of Solution	[309]
YmalDB: Exploring relational databases via result-driven recommendations	2013	Automatic Exploration	Proposal of Solution	[310]
ZoomTree: Unrestricted zoom paths in multiscale visual analysis of relational databases	2011	Visual Optimizations	Proposal of Solution	[311]
dbTouch: Analytics at your Fingertips.	2013	Novel Query Interfaces	Proposal of Solution	[312]
iOLAP: Managing Uncertainty for Efficient Incremental OLAP	2016	Query Approximation	Proposal of Solution	[313]

without replacement, so in each iteration new data points are taken into account, improving the selectivity of the method.

D. SUMMARY

We can see some commonalities looking at the underlying techniques used by the solutions described above:

First, for providing access to raw files, code generation and positional mapping seem to provide a good solution. Both are implemented either directly - PostgresRaw - or used via integration with an existing implementation - DiNoDB. Isolating the raw data access as a database operator composes well for all studied solutions regardless of the framework of reference - workflow, PostgreSQL or SciDB.

Second, to provide the interactivity on a distributed system, the engine needs to approximate the results using a deadline or an accuracy requirement as a stop condition. The resiliency and the low latency are achieved by being capable of processing only parts of the data, via sampling - BlinkDB -, pre-computed summaries - AccuracyTrader - or both. In either case, error estimation becomes an important part of the system, both internally and as part of the interface exposed to the user.

VI. CONCLUSIONS

In this systematic mapping study we have detailed the method that we followed to gather and filter papers related to *data exploration*, searching for solutions that tackle big data volumes, stored in a distributed way and with a low latency. This process have produced 242 papers, which we have classified according to their approach [8] on one axis, and to their research type [41] on another.

The results suggest that plenty of solutions have been proposed by researchers. However, there is rarely any follow up, at least published, on their practical implementation, be it to confirm a successful introduction to users or to evaluate other tools already in place. Unfortunately, this is not different to the state of other areas of the computing sciences.

We have found evidence that code generation is a well-proven approach for accessing raw data files, although most solutions have not been generalized onto a distributed environment.

AQP research can bring response times down to a latency suitable for interactive exploration. However, the overlap between raw data files and approximate query processing still seems to be an area where more research may be needed.

In general, there are building blocks that satisfy each one of the three requirements that we want to satisfy, albeit separately. However, it is unclear how difficult it would be to integrate or implement them in practice.

Finally, it seems likely that the future will bring even more powerful building blocks for data exploration, resulting in flexible systems tailored to specific needs and capable of adapting themselves to changes on the workflow, allowing users to focus on the information rather than on how to treat performantly the raw data.

APPENDIX
RESULTS OF THE MAPPING STUDY

See Tables.

REFERENCES

[1] G. Piatetsky-Shapiro, "Knowledge discovery in real databases: A report on the IJCAI-89 workshop," *AI Mag.*, vol. 11, no. 5, pp. 68–70, 1991. [Online]. Available: http://dl.acm.org/citation.cfm?id=124898.124915

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Towards a unifying framework," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1996, pp. 82–88. [Online]. Available: http://dl.acm.org/citation.cfm?id=3001460.3001477

[3] T. Reinartz, *Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-World Domains*. Berlin, Germany: Springer-Verlag, 1999.

[4] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *J. Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[5] D. Geer and J. Jacobs, "Exploring with a purpose," *Login*, vol. 39, pp. 47–49, Jun. 2014.

[6] R. B. Miller, "Response time in man-computer conversational transactions," in *Proc. Fall Joint Comput. Conf., Part I (AFIPS)*, Dec. 1968, pp. 267–277, doi: 10.1145/1476589.1476628.

[7] M. Kersten, S. Idreos, S. Manegold, and E. Liarou, "The researcher's guide to the data deluge: Querying a scientific database in just a few seconds," *Proc. VLDB Endowment*, vol. 3, no. 3, p. 1474, 2011.

[8] S. Idreos, O. Papaemmanouil, and S. Chaudhuri, "Overview of data exploration techniques," in *Proc. ACM SIGMOD Int. Conf. Manage. Data-SIGMOD*, 2015, pp. 277–281. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2723372.2731084

[9] M. Karpathiotakis, M. Branco, I. Alagiannis, and A. Ailamaki, "Adaptive query processing on RAW data," *Proc. VLDB Endowment*, vol. 7, no. 12, pp. 1119–1130, 2014-08. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2732977.2732986

[10] M. Karpathiotakis, I. Alagiannis, and A. Ailamaki, "Fast queries over heterogeneous data through engine customization," *Proc. VLDB Endowment*, vol. 9, no. 12, pp. 972–983, 2016. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2994509.2994516, doi: 10.14778/2994509.2994516.

[11] V. Silva, D. de Oliveira, P. Valduriez, and M. Mattoso, "Analyzing related raw data files through dataflows," *Concurrency Comput., Pract. Exper.*, vol. 28, no. 8, pp. 2528–2545, 2016, doi: 10.1002/cpe.3616.

[12] J. P. Baud *et al.*, "The LHCb data management system," *J. Phys., Conf. Ser.*, vol. 396, no. 3, p. 32023, 2012. [Online]. Available: http://stacks.iop.org/1742-6596/396/i=3/a=032023

[13] I. Bird, "WLCG: Report on project status, resources and financial plan," CERN, Geneva, Switzerland, Tech. Rep. CERN-RRB-2016-123, 2016.

[14] H. Rousseau. (2018). *EOS Ops at CERN*. [Online]. Available: https://indico.cern.ch/event/656157/contributions/2866314/

[15] M. Lamanna, "Large-scale data services for science: Present and future challenges," *Phys. Particles Nuclei Lett.*, vol. 13, no. 5, pp. 676–680, 2016, doi: 10.1134/S1547477116050344.

[16] S. Idreos, I. Alagiannis, R. Johnson, and A. Ailamaki, "Here are my data files. Here are my queries. Where are my results?" in *Proc. 5th Biennial Conf. Innov. Data Syst. Res. (CIDR)*, 2011, pp. 57–68.

[17] I. Alagiannis, R. Borovica-Gajic, M. Branco, S. Idreos, and A. Ailamaki, "NoDB: Efficient query execution on raw data files," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2012, vol. 58, no. 12, pp. 112–121. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2847579.2830508

[18] Y. Cheng and F. Rusu, "SCANRAW: A database meta-operator for parallel in-situ processing and loading," *ACM Trans. Database Syst.*, vol. 40, no. 3, pp. 1–45, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2838914.2818181

[19] A. Anagnostou, M. Olma, and A. Ailamaki, "Alpine: Efficient *in-situ* data exploration in the presence of updates," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 1651–1654. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3035918.3058743

[20] M. Olma, M. Karpathiotakis, I. Alagiannis, M. Athanassoulis, and A. Ailamaki, "Slalom: Coasting through raw data via adaptive partitioning and indexing," *Proc. VLDB Endowment*, vol. 10, no. 10, pp. 1106–1117, 2017. [Online]. Available: https://dl.acm.org/citation.cfm?id=3115415

[21] F. Olken and D. Rotem, "Simple random sampling from relational databases," in *Proc. 12th Int. Conf. Very Large Data Bases (VLDB)*. San Mateo, CA, USA: Morgan Kaufmann, 1986, pp. 160–169. [Online]. Available: http://dl.acm.org/citation.cfm?id=645913.671474

[22] S. Agarwal *et al.*, "Knowing when you're wrong: Building fast and reliable approximate query processing systems," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 481–492. [Online]. Available: http://dl.acm.org/citation.cfm?id=2593667

[23] S. Chaudhuri, B. Ding, and S. Kandula, "Approximate query processing: No silver bullet," in *Proc. ACM Int. Conf. Manage. Data*, 2017, pp. 511–519. [Online]. Available: https://dl.acm.org/citation.cfm?id=3056097

[24] D. Moritz and D. Fisher, "What users don't expect about exploratory data analysis on approximate query processing systems," in *Proc. 2nd Workshop Hum.-Loop Data Anal. (HILDA)*, 2017, pp. 9:1–9:4, doi: 10.1145/3077257.3077258.

[25] K. Zeng, S. Gao, B. Mozafari, and C. Zaniolo, "The analytical bootstrap: A new method for fast error estimation in approximate query processing," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2014, pp. 277–288, doi: 10.1145/2588555.2588579.

[26] Y. Wang, X. Xu, Y. Xia, and Q. Fang, "AQP++: A hybrid approximate query processing framework for generalized aggregation queries," in *Proc. Int. Conf. Adv. Cloud Big Data (CBD)*, 2017, pp. 56–62. [Online]. Available: http://ieeexplore.ieee.org/document/7815186/

[27] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*. New York, NY, USA: Springer, 1992, pp. 569–593, doi: 10.1007/978-1-4612-4380-9%7B%5C_%7D41.

[28] A. Galakatos, A. Crotty, E. Zgraggen, C. Binnig, and T. Kraska, "Revisiting Reuse for Approximate Query processing," *Proc. VLDB Endowment*, vol. 10, no. 1, pp. 1142–1153, 2017. [Online]. Available: http://www.vldb.org/pvldb/vol10/p1142-galakatos.pdf

[29] N. Potti and J. M. Patel, "DAQ: A new paradigm for approximate query processing," *Proc. VLDB Endowment*, vol. 8, no. 9, pp. 898–909, 2015, doi: 10.14778/2777598.2777599.

[30] Y. Park, A. S. Tajik, M. Cafarella, and B. Mozafari, "Database learning: Toward a database that becomes smarter every time," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 587–602. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3035918.3064013

[31] B. Mozafari, "Approximate query engines: Commercial challenges and research opportunities," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 5–8. [Online]. Available: https://dl.acm.org/citation.cfm?id=3056098

[32] P. Yongjoo, "Active database learning," in *Proc. CIDR*, 2017, p. 2. [Online]. Available: http://cidrdb.org/cidr2017/gongshow/abstracts/cidr2017%7B%5C_%7D54.pdf

[33] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin, Madison, WI, USA, Tech. Rep. 1648, 2010. [Online]. Available: http://burrsettles.com/pub/settles.activelearning.pdf

[34] D. Narayanan, A. Donnelly, R. Mortier, and A. Rowstron, "Delay aware querying with Seaweed," *VLDB J.*, vol. 17, no. 2, pp. 315–331, 2008. [Online]. Available: http://link.springer.com/article/10.1007/s00778-007-0060-3

[35] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," *Engineering*, vol. 2, p. 1051, Jan. 2007.

[36] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proc. 12th Int. Conf. Eval. Assessment Softw. Eng.*, vol. 17, 2007, p. 10. [Online]. Available: http://dl.acm.org/citation.cfm?id=2227115.2227123

[37] R. Brun and F. Rademakers, "ROOT—An object oriented data analysis framework," *Nucl. Instrum. Methods Phys. Res. A, Accel. Spectrom. Detect. Assoc. Equip.*, vol. 389, nos. 1–2, pp. 81–86, 1997.

[38] A. J. P. Janyst and L. Janyst, "Exabyte scale storage at CERN," *J. Phys., Conf. Ser.*, vol. 331, no. 5, p. 52015, 2011. [Online]. Available: http://stacks.iop.org/1742-6596/331/i=5/a=052015

[39] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quart.*, vol. 26, no. 2, pp. 23–33, 2002.

[40] S. Jalali and C. Wohlin, "Systematic literature studies: Database searches vs. backward snowballing," in *Proc. ACM-IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Sep. 2012, pp. 29–38. [Online]. Available: http://dl.acm.org/citation.cfm?id=2372257

[41] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion," *Requirements Eng.*, vol. 11, no. 1, pp. 102–107, 2006. [Online]. Available: https://dl.acm.org/citation.cfm?id=1107683

[42] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Inf. Softw. Technol.*, vol. 55, no. 12, pp. 2049–2075, 2013.

[43] M. Jorgensen and M. Shepperd, "A systematic review of software development cost estimation studies," *IEEE Trans. Softw. Eng.*, vol. 33, no. 1, pp. 33–53, Jan. 2007.

[44] J. Bailey, C. Zhang, D. Budgen, M. Turner, and S. Charters, "Search engine overlaps: Do they agree or disagree?" in *Proc. Workshops (ICSE), 2nd Int. Workshop Realising Evidence-Based Softw. Eng. (REBSE)*, May 2007, p. 2.

[45] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, 2007.

[46] A. Oakley, "Research evidence, knowledge management and educational practice: Early lessons from a systematic approach," *London Rev. Edu.*, vol. 1, no. 1, pp. 21–33, 2003. [Online]. Available: http://www.ingentaconnect.com/contentone/ioep/clre/2003/00000001/00000001/art00004

[47] A. Siddiqa, A. Karim, T. Saba, and V. Chang, "On the analysis of big data indexing execution strategies," *J. Intell. Fuzzy Syst.*, vol. 32, no. 5, pp. 3259–3271, 2017. [Online]. Available: https://eprints.soton.ac.uk/399946/

[48] K. Zoumpatianos, "Query workloads for data series indexes," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1603–1612. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2783258.2783382

[49] T. Palpanas, ''Data series management: The road to big sequence analytics,'' *SIGMOD Rec.*, vol. 44, no. 2, pp. 47–52, 2015. [Online]. Available: https://dl.acm.org/citation.cfm?id=2814719
[50] A. Lawrence, J. Thomas, J. Houghton, and P. Weldon, ''Collecting the evidence: Improving access to grey literature and data for public policy and practice,'' *Austral. Acad. Res. Libraries*, vol. 46, no. 4, pp. 229–249, 2015, doi: 10.1080/00048623.2015.1081712.
[51] W. F. Tichy, P. Lukowicz, L. Prechelt, and E. A. Heinz, ''Experimental evaluation in computer science: A quantitative study,'' *J. Syst. Softw.*, vol. 28, no. 1, pp. 9–18, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016412129400111Y
[52] M. V. Zelkowitz and D. Wallace, ''Experimental validation in software engineering,'' *Inf. Softw. Technol.*, vol. 39, no. 11, pp. 735–743, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950584997000256
[53] D. I. Sjøberg *et al.*, ''A survey of controlled experiments in software engineering,'' *IEEE Trans. Softw. Eng.*, vol. 31, no. 9, pp. 733–753, Sep. 2005.
[54] D. Han, Y.-M. Nam, and M.-S. Kim, ''A distributed *in-situ* analysis method for large-scale scientific data,'' in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 69–75. [Online]. Available: http://ieeexplore.ieee.org/document/7881718/
[55] M. Stonebraker and U. Cetintemel, ''One size fits all: An idea whose time has come and gone,'' in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, 2005, pp. 2–11.
[56] D. Budgen, M. Turner, P. Brereton, and B. Kitchenham, ''Using mapping studies in software engineering,'' in *Proc. PPIG*, vol. 2, 2008, pp. 195–204. [Online]. Available: http://www.ppig.org/papers/20th-budgen.pdf%7B%5C%%7D5Cnhttp://www.inf.puc-rio.br/%7B~%7Dinf2921/2014%7B%5C_%7D2/docs/artigos/Using%20Mapping%20Studies%20in%20Software%20Engineering.pdf
[57] M. MacLure, '''Clarity bordering on stupidity': Where's the quality in systematic review?'' *J. Edu. Policy*, vol. 20, no. 4, pp. 393–416, 2005. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02680930500131801
[58] M. Stonebraker, P. Brown, A. Poliakov, and S. Raman, *The Architecture of SciDB*. Berlin, Germany: Springer, 2011, pp. 1–16. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-22351-8%7B%5C_%7D1
[59] *HDF Group*. Accessed: May 15, 2018. [Online]. Available: https://www.hdfgroup.org/
[60] Y. Tian, I. Alagiannis, E. Liarou, A. Ailamaki, P. Michiardi, and M. Vukolić, ''DiNoDB: Efficient large-scale raw data analytics,'' in *Proc. 1st Int. Workshop Bringing Value 'Big Data' Users (Data4U)*, 2014, pp. 1–6. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2658840.2658841
[61] V. Silva *et al.*, ''Raw data queries during data-intensive parallel workflow execution,'' *Future Gener. Comput. Syst.*, vol. 75, pp. 402–422, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17300237
[62] E. Ogasawara, D. De Oliveira, P. Valduriez, J. Dias, F. Porto, and M. Mattoso, ''An algebraic approach for data-centric scientific workflows,'' *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 1328–1339, 2011.
[63] K. Wu *et al.*, ''FastBit: Interactively searching massive data,'' *J. Phys., Conf. Ser.*, vol. 180, no. 1, p. 012053, 2009.
[64] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica, ''BlinkDB: queries with bounded errors and bounded response times on very large data,'' in *Proc. 8th ACM Eur. Conf. Comput. Syst. (EuroSys)*, 2013, p. 29. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2465351.2465355
[65] S. Lohr, *Sampling: Design and Analysis*. Toronto ON, Canada: Nelson Education, 2009.
[66] L. Battle, M. Stonebraker, and R. Chang, ''Dynamic reduction of query result sets for interactive visualizaton,'' in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/6691708/
[67] N. Kamat, P. Jayachandran, K. Tunga, and A. Nandi, ''Distributed and interactive cube exploration,'' in *Proc. IEEE 30th Int. Conf. Data Eng.*, Mar. 2014, pp. 472–483. [Online]. Available: http://ieeexplore.ieee.org/document/6816674/
[68] J. Gray *et al.*, ''Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals,'' *Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 29–53, 1997.
[69] R. Han, S. Huang, F. Tang, F. Chang, and J. Zhan, ''Accuracy-Trader: Accuracy-aware approximate processing for low tail latency and high result accuracy in cloud online services,'' in *Proc. 45th Int. Conf. Parallel Process.*, vol. 8, 2016, pp. 278–287. [Online]. Available: https://arxiv.org/abs/1607.02734
[70] S.-S. Kim, T. Lee, M. Chung, and J. Won, ''Flying KIWI: Design of approximate query processing engine for interactive data analytics at scale,'' in *Proc. Int. Conf. Big Data Appl. Services (BigDAS)*, 2015, pp. 206–207, doi: 10.1145/2837060.2837096.
[71] Y. Wang, L. Chen, and G. Agrawal, ''Supporting online analytics with user-defined estimation and early termination in a MapReduce-like framework,'' in *Proc. Int. Workshop Data-Intensive Scalable Comput. Syst. (DISCS)*, 2015, pp. 1–8. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2831244.2831247
[72] J. Moehrmann, A. Burkovski, E. Baranovskiy, G. Heinze, A. Rapoport, and G. Heidemann, ''A discussion on visual interactive data exploration using self-organizing maps,'' in in *Proc. Int. Workshop Self-Org. Maps*, 2011, pp. 178–187. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-21566-7%7B%5C_%7D18
[73] P. Liakos, P. Koltsida, G. Kakaletris, P. Baumann, Y. Ioannidis, and A. Delis, ''A distributed infrastructure for earth-science big data retrieval,'' *Int. J. Cooperat. Inf. Syst.*, vol. 24, no. 2, p. 1550002, 2015. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0218843015500021
[74] H. Doraiswamy, H. T. Vo, C. T. Silva, and J. Freire, ''A GPU-based index to support interactive spatio-temporal queries over historical data,'' in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 1086–1097. [Online]. Available: http://ieeexplore.ieee.org/document/7498315/
[75] J. Aligon, K. Boulil, P. Marcel, and V. Peralta, ''A holistic approach to OLAP sessions composition: The falseto experience,'' in *Proc. 17th Int. Workshop Data Warehousing OLAP (DOLAP)*, 2014, pp. 37–46. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2666158.2666179
[76] H. Liu, J. X. Yu, J. Zeleznikow, and Y. Guan, ''A logic-based approach to mining inductive databases,'' in *Proc. Int. Conf. Comput. Sci.*, 2007, pp. 270–277. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-72584-8%7B%5C_%7D35
[77] M. Brucato, A. Abouzied, and A. Meliou, ''A scalable execution engine for package queries,'' *ACM SIGMOD Rec.*, vol. 46, no. 1, pp. 24–31, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3093754.3093761
[78] D. Nicklas, T. Schwarz, and B. Mitschang, ''A schema-based approach to enable data integration on the fly,'' *Int. J. Cooperat. Inf. Syst.*, vol. 26, no. 1, p. 1650010, 2017. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0218843016500106
[79] A. Termehchy and B. Touri, ''A Signaling Game Approach to Databases Querying and Interaction,'' in *Proc. Int. Conf. Theory Inf. Retr.*, vol. 1644, 2016, pp. 361–364. [Online]. Available: https://dl.acm.org/citation.cfm?id=2809487
[80] N. Kamat and A. Nandi, ''A unified correlation-based approach to sampling over joins,'' in *Proc. 29th Int. Conf. Sci. Stat. Database Manage. (SSDBM)*, 2017, pp. 1–12. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3085504.3085524
[81] D. Han, Y.-M. Nam, and M.-S. Kim, ''A distributed in-situ analysis method for large-scale scientific data,'' in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 69–75. [Online]. Available: http://ieeexplore.ieee.org/document/7881718/
[82] M. S. Islam, C. Liu, and R. Zhou, ''A framework for query refinement with user feedback,'' *J. Syst. Softw.*, vol. 86, no. 6, pp. 1580–1595, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121213000265
[83] J. Zaia and J. A. L. Bernardes, ''A graphical system for interactive creation and exploration of dynamic information visualization,'' in *Proc. Int. Conf. Hum. Interface Manage. Inf.*, vol. 9734, 2016, pp. 214–225. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-40349-6%7B%5C_%7D21
[84] N. Bikakis, G. Papastefanatos, M. Skourla, and T. Sellis, ''A hierarchical aggregation framework for efficient multilevel visual exploration and analysis,'' *Semantic Web*, vol. 8, no. 1, pp. 139–179, 2017. [Online]. Available: https://arxiv.org/abs/1511.04750
[85] Y. Chen *et al.*, ''A study of SQL-on-Hadoop systems,'' in *Proc. Big Data Benchmarks, Perform. Optim., Emerg. Hardw.*, vol. 8807, 2014, pp. 154–166. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-13021-7%7B%5C_%7D12

[86] D. Taniar and W. Rahayu, "A taxonomy for region queries in spatial databases," *J. Comput. Syst. Sci.*, vol. 81, no. 8, pp. 1508–1531, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022000014001895

[87] F. Eichinger, P. Efros, S. Karnouskos, and K. Böhm, "A time-series compression technique and its application to the smart grid," *VLDB J. Int. J. Very Large Data Bases*, vol. 24, no. 2, pp. 193–218, 2015. [Online]. Available: http://link.springer.com/10.1007/s00778-014-0368-8

[88] K. Zoumpatianos, S. Idreos, and T. Palpanas, "ADS: The adaptive data series index," *VLDB J.*, vol. 25, no. 6, pp. 843–866, 2016. [Online]. Available: http://link.springer.com/article/10.1007/s00778-016-0442-5

[89] K. Dimitriadou, O. Papaemmanouil, and Y. Diao, "AIDE: An active learning-based approach for interactive data exploration," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 2842–2856, Nov. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7539596/

[90] S. Schuh and J. Dittrich, "AIR: Adaptive index replacement in Hadoop," in *Proc. 31st IEEE Int. Conf. Data Eng. Workshops*, Apr. 2015, pp. 22–29. [Online]. Available: http://ieeexplore.ieee.org/document/7129539/

[91] Y. Wang, X. Xu, Y. Xia, and Q. Fang, "AQP++: A hybrid approximate query processing framework for generalized aggregation queries," in *Proc. Int. Conf. Adv. Cloud Big Data (CBD)*, Aug. 2017, pp. 56–62. [Online]. Available: http://ieeexplore.ieee.org/document/7815186/

[92] Z. Hong, C. Yu, J. Wang, J. Xiao, C. Cui, and J. Sun, "AQUAdexIM: Highly efficient in-memory indexing and querying of astronomy time series images," *Exp. Astron.*, vol. 42, no. 3, pp. 387–405, 2016. [Online]. Available: https://link.springer.com/article/10.1007/s10686-016-9515-0

[93] M. Sassi, A. G. Touzi, H. Ounelli, and I. Aissa, "About database summarization," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 18, no. 2, pp. 133–151, 2010. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0218488510006453

[94] C. Koch, "Abstraction without regret in database systems building: A manifesto," *IEEE Data Eng. Bull.*, vol. 37, no. 1, pp. 70–79, 2014. [Online]. Available: http://infoscience.epfl.ch/record/197359

[95] M. S. Kester, M. Athanassoulis, and S. Idreos, "Access path selection in main-memory optimized data systems: Should i scan or should i probe?" in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 715–730. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3035918.3064049

[96] R. Han, S. Huang, F. Tang, F. Chang, and J. Zhan, "AccuracyTrader: Accuracy-aware approximate processing for low tail latency and high result accuracy in cloud online services," in *Proc. 45th Int. Conf. Parallel Process.*, no. 8, 2016, pp. 278–287. [Online]. Available: https://arxiv.org/abs/1607.02734

[97] P. Karras, A. Nikitin, M. Saad, R. Bhatt, D. Antyukhov, and S. Idreos, "Adaptive indexing over encrypted numeric data," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 171–183. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2882903.2882932

[98] H. Liu and Z. Liu, "Adaptive indexing approach for main memory column store," *J. Eng.*, vol. 2017, no. 2, pp. 26–32, 2016. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/joe.2016.0068

[99] M. Karpathiotakis, M. Branco, I. Alagiannis, and A. Ailamaki, "Adaptive query processing on RAW data," *Proc. VLDB Endowment*, vol. 7, no. 12, pp. 1119–1130, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2732977.2732986

[100] F. N. Afrati, P. V. Lekeas, and C. Li, "Adaptive-sampling algorithms for answering aggregation queries on Web sites," *Data Knowl. Eng.*, vol. 64, no. 2, pp. 462–490, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169023X07001814

[101] A. Anagnostou, M. Olma, and A. Ailamaki, "Alpine: Efficient *in-situ* data exploration in the presence of updates," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 1651–1654. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3035918.3058743

[102] C. Guo, Z. Wu, Z. He, and X. S. Wang, "An adaptive data partitioning scheme for accelerating exploratory spark SQL queries," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2017, pp. 114–128. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-55753-3%7B%5C_%7D8

[103] W. Liu, Y. Diao, and A. Liu, "An analysis of query-agnostic sampling for interactive data exploration," *Commun. Statist. Theory Methods*, vol. 47, no. 16, pp. 3820–3837, 2017, doi: 10.1080/03610926.2017.1363231.

[104] K. Kolomvatsos, C. Anagnostopoulos, and S. Hadjiefthymiades, "An efficient time optimized scheme for progressive analytics in big data," *Big Data Res.*, vol. 2, no. 4, pp. 155–165, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214579615000106

[105] H. J. Schulz, M. Angelini, G. Santucci, and H. Schumann, "An enhanced visualization process model for incremental visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 7, pp. 1830–1842, Jul. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7172541/

[106] F. M. Schuhknecht, A. Jindal, and J. Dittrich, "An experimental evaluation and analysis of database cracking," *VLDB J.*, vol. 25, no. 1, pp. 27–52, 2016. [Online]. Available: https://doi.org/10.1007/s00778-015-0397-y

[107] K. Kolomvatsos, "An intelligent, uncertainty driven aggregation scheme for streams of ordered sets," *Appl. Intell.*, vol. 45, no. 3, pp. 713–735, 2016. [Online]. Available: https://doi.org/10.1007/s10489-016-0789-8

[108] L. Braun *et al.*, "Analytics in motion: High performance event-processing and real-time analytics in the same database," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2015, pp. 251–264. [Online]. Available: http://doi.acm.org/10.1145/2723372.2742783

[109] N. Franciscus, X. Ren, and B. Stantic, "Answering temporal analytic queries over big data based on precomputing architecture," in *Proc. Intell. Inf. Database Syst.*, 2017, pp. 281–290. [Online]. Available: http://link.springer.com/10.1007/978-3-319-54472-4%7B%5C_%7D27

[110] S. A. Shaikh and H. Kitagawa, "Approximate OLAP on sustained data streams," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2017, pp. 102–118. [Online]. Available: http://link.springer.com/10.1007/978-3-319-55699-4%7B%5C_%7D7

[111] B. Mozafari, "Approximate query engines: Commercial challenges and research opportunities," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 5–8. [Online]. Available: https://dl.acm.org/citation.cfm?id=3056098

[112] S. Chaudhuri, B. Ding, and S. Kandula, "Approximate query processing: No silver bullet," *Proc. ACM Int. Conf. Manage. Data*, 2017, pp. 511–519. [Online]. Available: https://dl.acm.org/citation.cfm?id=3056097

[113] M. Streppel and K. Yi, "Approximate range searching in external memory," *Algorithmica*, vol. 59, no. 2, pp. 115–128, 2011. [Online]. Available: http://link.springer.com/article/10.1007/s00453-009-9297-0%20 and http://link.springer.com/chapter/10.1007/978-3-540-77120-3%7B%5C_%7D47

[114] P. Neophytou *et al.*, "AstroShelf: Understanding the universe through scalable navigation of a galaxy of annotations," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 713–716. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2213836.2213940

[115] M. Djedaini, P. Furtado, N. Labroche, P. Marcel, and V. Peralta, "Benchmarking exploratory OLAP," in *Performance Evaluation and Benchmarking. Traditional-Big Data-Internet Things* (Lecture Notes in Computer Science), vol. 10080. 2017, pp. 61–77. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-54334-5%7B%5C_%7D5

[116] A. Camerra, J. Shieh, T. Palpanas, T. Rakthanmanon, and E. Keogh, "Beyond one billion time series: Indexing and mining very large time series collections with *i*SAX2+," *Knowl. Inf. Syst.*, vol. 39, no. 1, pp. 123–151, 2014. [Online]. Available: http://link.springer.com/article/10.1007/s10115-012-0606-6

[117] S. L. Xi, O. Babarinsa, M. Athanassoulis, and S. Idreos, "Beyond the wall: Near-data processing for databases," in *Proc. 11th Int. Workshop Data Manage. New Hardw. (DaMoN)*, 2015, pp. 2:1–2:10. [Online]. Available: http://doi.acm.org/10.1145/2771937.2771945

[118] Y. Cheng, W. Zhao, and F. Rusu, "Bi-level online aggregation on raw data," in *Proc. 29th Int. Conf. Sci. Stat. Database Manage. (SSDBM)*, 2017, pp. 1–12. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3085504.3085514

[119] T. Palpanas, "Big sequence management: A glimpse of the past, the present, and the future," in *Proc. Int. Conf. Current Trends Theory Pract. Inform.*, 2016, pp. 63–80. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-662-49192-8%7B%5C_%7D6

[120] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica, "BlinkDB: Queries with bounded errors and bounded response times on very large data," in *Proc. 8th ACM Eur. Conf. Comput. Syst. (EuroSys)*, 2013, p. 29. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2465351.2465355

[121] J. Arulraj, A. Pavlo, and P. Menon, "Bridging the archipelago between row-stores and column-stores for hybrid workloads," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2016, pp. 583–598. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2882903.2915231
[122] Y. Klonatos, C. Koch, T. Rompf, and H. Chafi, "Building efficient query engines in a high-level language," *Proc. VLDB Endowment*, vol. 7, no. 10, pp. 853–864, 2014. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2732951.2732959
[123] P. Thanisch, J. Nummenmaa, T. Niemi, and M. Niinim'aki, "Cell-at-a-time approach to lazy evaluation of dimensional aggregations," in *Proc. Int. Conf. Data Warehousing Knowl. Discovery*, 2013, pp. 349–358. [Online]. Available: http://link.springer.com/10.1007/978-3-642-40131-2%7B%5C_%7D31
[124] S. Chen, "Cheetah: A high performance, custom data warehouse on top of MapReduce," *Proc. VLDB Endowment*, vol. 3, nos. 1–2, pp. 1459–1468, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1920841.1921020
[125] B. Mozafari, E. Z. Y. Goh, and D. Y. Yoon, "CliffGuard: A principled framework for finding robust database designs," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2015, pp. 1167–1182. [Online]. Available: http://doi.acm.org/10.1145/2723372.2749454
[126] T. Sellam and M. Kersten, "Cluster-driven navigation of the query space," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1118–1131, May 2016. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7374727/
[127] C. Demiralp. (2016). *Clustrophile: A Tool for Visual Clustering Analysis*. [Online]. Available: http://poloclub.gatech.edu/idea2016/papers/p37-demiralp.pdf
[128] E. Wu *et al.*, "Combining design and performance in a data visualization management system," in *Proc. CIDR*, 2017, pp. 1–11.
[129] A. Cheung and A. Solar-Lezama, "Computer-assisted query formulation," *Found. Trends Signal Process.*, vol. 8, nos. 1–2, pp. 1–126, 2016. [Online]. Available: https://homes.cs.washington.edu/%7B~%7Dakcheung/papers/fntSurvey15.pdf
[130] G. Graefe, F. Halim, S. Idreos, H. Kuno, and S. Manegold, "Concurrency control for adaptive indexing," *Proc. VLDB Endowment*, vol. 5, no. 7, pp. 656–667, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2180912.2180918
[131] Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska, "Controlling false discoveries during interactive data exploration," in *Proc. ACM Int. Conf. Manage. Data*, 2017, pp. 527–540. [Online]. Available: http://arxiv.org/abs/1612.01040%20 and https://dl.acm.org/citation.cfm?id=3064019
[132] Y. Zhuang *et al.*, "D-Ocean: An unstructured data management system for data ocean environment," *Frontiers Comput. Sci.*, vol. 10, no. 2, pp. 353–369, 2016. [Online]. Available: http://link.springer.com/article/10.1007/s11704-015-5045-6
[133] N. Potti and J. M. Patel, "DAQ: A new paradigm for approximate query processing," *Proc. VLDB Endowment*, vol. 8, no. 9, pp. 898–909, 2015, doi: 10.14778/2777598.2777599.
[134] A. Dziedzic, M. Karpathiotakis, I. Alagiannis, R. Appuswamy, and A. Ailamaki, "DBMS data loading: An analysis on modern hardware," in *Data Management on New Hardware*, 2017. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-56111-0%7B%5C_%7D6
[135] S. Lakshminarasimhan *et al.*, "DIRAQ: Scalable *in situ* data- and resource-aware indexing for optimized query performance," *Cluster Comput.*, vol. 17, no. 4, pp. 1101–1119, 2014. [Online]. Available: http://link.springer.com/article/10.1007/s10586-014-0358-z
[136] A. Wasay, X. Wei, N. Dayan, and S. Idreos, "Data canopy: Accelerating exploratory statistical analysis," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 557–572. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3035918.3064051
[137] J. Cumin, J.-M. Petit, V.-M. Scuturici, and S. Surdu, "Data exploration with SQL using machine learning techniques," in *Proc. Int. Conf. Extending Database Technol. (EDBT)*, 2017, pp. 96–107. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01455715
[138] M. Khan, L. Xu, A. Nandi, and J. M. Hellerstein, "Data tweening: Incremental visualization of data transforms," *Proc. VLDB Endowment*, vol. 10, no. 6, pp. 661–672, 2017, doi: 10.14778/3055330.3055333.
[139] T. Palpanas, "Data series management: The road to big sequence analytics," *ACM SIGMOD Rec.*, vol. 44, no. 2, pp. 47–52, 2015. [Online]. Available: https://dl.acm.org/citation.cfm?id=2814719
[140] M. Ivanova, M. Kersten, and S. Manegold, "Data vaults: A symbiosis between database technology and scientific file repositories," in *Scientific Statistical Database Management* (Lecture Notes in Computer Science), vol. 7338. 2012, pp. 485–494. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-31235-9%7B%5C_%7D32
[141] A. Abouzied, J. Hellerstein, and A. Silberschatz, "DataPlay: Interactive tweaking and example-driven correction of graphical database queries," in *Proc. 25th Annu. ACM Symp. User Interface Softw. Technol.*, 2012, pp. 207–218. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2380116.2380144
[142] H. Pirk, E. Petraki, S. Idreos, S. Manegold, and M. Kersten, "Database cracking: Fancy scan, not poor man's sort!" in *Proc. 10th Int. Workshop Data Manage. New Hardw. (DaMoN)*, 2014, pp. 4:1–4:8. [Online]. Available: http://doi.acm.org/10.1145/2619228.2619232
[143] Y. Park, A. S. Tajik, M. Cafarella, and B. Mozafari, "Database learning: Toward a database that becomes smarter every time," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 587–602. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3035918.3064013
[144] D. Narayanan, A. Donnelly, R. Mortier, and A. Rowstron, "Delay aware querying with Seaweed," *VLDB J.*, vol. 17, no. 2, pp. 315–331, 2008. [Online]. Available: http://link.springer.com/article/10.1007/s00778-007-0060-3
[145] R. Huang, R. Chirkova, and Y. Fathi, "Deterministic view selection for data-analysis queries: Properties and algorithms," in *Proc. East Eur. Conf. Adv. Databases Inf. Syst.*, 2012, pp. 195–208. [Online]. Available: http://link.springer.com/10.1007/978-3-642-33074-2%7B%5C_%7D15
[146] Y. Tian, I. Alagiannis, E. Liarou, A. Ailamaki, P. Michiardi, and M. Vukolić, "DiNoDB: Efficient large-scale raw data analytics," in *Proc. 1st Int. Workshop Bringing Value Big Data Users (Data4U)*, 2014, pp. 1–6. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2658840.2658841
[147] Y. Shen, K. Chakrabarti, S. Chaudhuri, B. Ding, and L. Novik, "Discovering queries based on example tuples," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2014, pp. 493–504. [Online]. Available: http://doi.acm.org/10.1145/2588555.2593664
[148] N. Kamat, P. Jayachandran, K. Tunga, and A. Nandi, "Distributed and interactive cube exploration," in *Proc. IEEE 30th Int. Conf. Data Eng.*, Mar. 2014, pp. 472–483. [Online]. Available: http://ieeexplore.ieee.org/document/6816674/
[149] H. A. Khan, M. A. Sharaf, and A. Albarrak, "DivIDE: Efficient diversification for interactive data exploration," in *Proc. 26th Int. Conf. Sci. Stat. Database Manage. (SSDBM)*, 2014, pp. 15:1–15:12. [Online]. Available: http://doi.acm.org/10.1145/2618243.2618253
[150] Z. Hussain, H. A. Khan, and M. A. Sharaf, "Diversifying with few regrets, but too few to mention," in *Proc. 2nd Int. Workshop Explor. Search Databases Web (ExploreDB)*, 2015, pp. 27–32. [Online]. Available: http://doi.acm.org/10.1145/2795218.2795225
[151] A. Bampoulidis, J. A. Palotti, M. Lupu, J. Brassey, and A. Hanbury, "Does online evaluation correspond to offline evaluation in query auto completion?" in *Proc. Eur. Conf. Inf. Retr.*, 2017, pp. 713–719. [Online]. Available: http://link.springer.com/10.1007/978-3-319-56608-5%7B%5C_%7D70
[152] L. Battle, R. Chang, and M. Stonebraker, "Dynamic prefetching of data tiles for interactive visualization," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2016, pp. 1363–1375. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2882903.2882919
[153] L. Battle, M. Stonebraker, and R. Chang, "Dynamic reduction of query result sets for interactive visualizaton," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/6691708/
[154] Y. Wu, B. Harb, J. Yang, and C. Yu, "Efficient evaluation of object-centric exploration queries for visualization," *Proc. VLDB Endowment*, vol. 8, no. 12, pp. 1752–1763, 2015. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2824032.2824072
[155] A. M. Albarrak and M. A. Sharaf, "Efficient schemes for similarity-aware refinement of aggregation queries," *World Wide Web*, vol. 20, no. 6, pp. 1237–1267, 2017. [Online]. Available: https://link.springer.com/article/10.1007/s11280-017-0434-4
[156] K. Pantazos, S. Lauesen, and R. Vatrapu, "End-user development of information visualization," in *End-User Development*. 2013, pp. 104–119. [Online]. Available: http://link.springer.com/10.1007/978-3-642-38706-7%7B%5C_%7D9

[157] A. Yasir, M. K. Swamy, P. K. Reddy, and S. Bhalla, "Enhanced query-by-object approach for information requirement elicitation in large databases," in *Big Data Analytics*, vol. 7678. 2012, pp. 26–41. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-35542-4

[158] H. Li, H. Li, M. Chen, Z. Dai, M. Zhu, and M. Huang, "Enhancing parallel data loading for large scale scientific database," in *Big Data Analytics*. 2015, pp. 149–162. [Online]. Available: http://link.springer.com/10.1007/978-3-319-27122-4

[159] L. Zhu, X. Song, and C. Liu, "Evaluating a stream of relational KNN queries by a knowledge base," *Int. J. Cooperat. Inf. Syst.*, vol. 24, no. 2, p. 1550003, 2015. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0218843015500033

[160] V. Niennattrakul, P. Ruengronghirunya, and C. A. Ratanamahatana, "Exact indexing for massive time series databases under time warping distance," *Data Mining Knowl. Discovery*, vol. 21, no. 3, pp. 509–541, 2010. [Online]. Available: http://link.springer.com/article/10.1007/s10618-010-0165-y

[161] D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas, "Exemplar queries: A new way of searching," *VLDB J.*, vol. 25, no. 6, pp. 741–765, 2016. [Online]. Available: https://link.springer.com/article/10.1007/s00778-016-0429-2

[162] K. Panev, S. Michel, E. Milchevski, and K. Pal, "Exploring databases via reverse engineering ranking queries with PALEO," *Proc. VLDB Endowment*, vol. 9, no. 13, pp. 1525–1528, 2016, doi: 10.14778/3007263.3007300.

[163] S. Wang, D. Maier, and B. C. Ooi, "Fast and adaptive indexing of multi-dimensional observational data," *Proc. VLDB Endowment*, vol. 9, no. 14, pp. 1683–1694, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3007328.3007334

[164] M. Karpathiotakis, I. Alagiannis, and A. Ailamaki, "Fast queries over heterogeneous data through engine customization," *Proc. VLDB Endowment*, vol. 9, no. 12, pp. 972–983, 2016. [Online]. Available: http://dias.epfl.ch/files/content/sites/dias/files/Student%20Projects/p972-karpathiotakis.pdf%20 and http://dl.acm.org/itation.cfm?doid=2994509.2994516

[165] T. Sellam and M. Kersten, "Fast, explainable view detection to characterize exploration queries," in *Proc. 28th Int. Conf. Sci. Stat. Database Manage. (SSDBM)*, 2016, pp. 1–12. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2949689.2949692

[166] T. Siddiqui *et al.*, "Fast-forwarding to desired visualizations with zenvisage," in *Proc. CIDR*, 2017.

[167] V. Le and S. Gulwani, "FlashExtract: A framework for data extraction by examples," in *Proc. 35th ACM SIGPLAN Conf. Program. Lang. Design Implement. (PLDI)*, 2013, vol. 49, no. 6, pp. 542–553. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2594291.2594333

[168] S.-S. Kim, T. Lee, M. Chung, and J. Won, "Flying KIWI: Design of approximate query processing engine for interactive data analytics at scale," in *Proc. Int. Conf. Big Data Appl. Services (BigDAS)*, 2015, pp. 206–207. [Online]. Available: http://doi.acm.org/10.1145/2837060.2837096

[169] A. Nandi, L. Jiang, and M. Mandel, "Gestural query specification," *Proc. VLDB Endowment*, vol. 7, no. 4, pp. 289–300, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2732240.2732247

[170] I. Alagiannis, S. Idreos, and A. Ailamaki, "H2O: A hands-free adaptive store," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 1103–1114. [Online]. Available: https://dl.acm.org/citation.cfm?id=2588555.2610502

[171] C. A. L. Pahins, S. A. Stephens, C. Scheidegger, and J. A. L. Comba, "Hashedcubes: Simple, low memory, real-time visual exploration of big data," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 671–680, Jan. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7539326/

[172] E. Petraki, S. Idreos, and S. Manegold, "Holistic indexing in main-memory column-stores," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1153–1166. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2723372.2723719

[173] E. Zgraggen, A. Galakatos, A. Crotty, J.-D. Fekete, and T. Kraska, "How progressive visualizations affect exploratory analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 8, pp. 1977–1987, Aug. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7563865/

[174] C. Xie, W. Zhong, J. Kong, W. Xu, K. Mueller, and F. Wang, "IEVQ: An iterative example-based visual query for pathology database," in *Proc. VLDB Workshop Data Manage. Anal. Med. Healthcare*, 2017, pp. 29–42. [Online]. Available: http://link.springer.com/10.1007/978-3-319-57741-8

[175] M. X. Bornschlegl *et al.*, "IVIS4BigData: A reference model for advanced visual interfaces supporting big data analysis in virtual research environments," in *Advanced Visual Interfaces. Supporting Big Data Applications* (Lecture Notes in Computer Science), vol. 10084. 2016, pp. 1–18. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-50070-6

[176] D. R. Krishnan, D. L. Quoc, P. Bhatotia, C. Fetzer, and R. Rodrigues, "IncApprox: A data analytics system for incremental approximate computing," in *Proc. 25th Int. Conf. World Wide Web (WWW)*, International World Wide Web Conferences Steering Committee, 2016, pp. 1133–1144, doi: 10.1145/2872427.2883026.

[177] K. Zoumpatianos, S. Idreos, and T. Palpanas, "Indexing for interactive exploration of big data series," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 1555–1566. [Online]. Available: https://dl.acm.org/citation.cfm?id=2610498

[178] S. I. Mohammed, F. A. Omara, and H. M. Sharaf, "Information retrieval using dynamic indexing," in *Proc. 9th Int. Conf. Inform. Syst.*, Dec. 2014, pp. PDC-93–PDC-101. [Online]. Available: http://ieeexplore.ieee.org/document/7036684/

[179] W. Liu, Y. Diao, and A. Liu. (2016). *Initial Sampling for Automatic Interactive Data Exploration*. [Online]. Available: http://people.cs.umass.edu/

[180] D. Ślęzak and M. Kowalski, "Intelligent data granulation on load: Improving infobright's knowledge grid," in *Proc. Int. Conf. Future Gener. Inf. Technol.*, 2009, pp. 12–25. [Online]. Available: http://link.springer.com/10.1007/978-3-642-10509-8

[181] M. Kahng, S. B. Navathe, J. T. Stasko, and D. H. P. Chau, "Interactive browsing and navigation in relational databases," *Proc. VLDB Endowment*, vol. 9, no. 12, pp. 1017–1028, 2016. [Online]. Available: http://arxiv.org/abs/1603.02371

[182] A. Kalinin, U. Cetintemel, and S. Zdonik, "Interactive data exploration using semantic windows," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 505–516, doi: 10.1145/2588555.2593666.

[183] A. Bonifati, R. Ciucanu, and S. Staworko, "Interactive inference of join queries," in *Proc. Int. Conf. Extending Database Technol. (EDBT)*, 2014, pp. 451–462. [Online]. Available: https://hal.inria.fr/hal-00875680

[184] J. Fan, G. Li, and L. Zhou, "Interactive SQL query suggestion: Making databases user-friendly," in *Proc. IEEE 27th Int. Conf. Data Eng.*, Apr. 2011, pp. 351–362. [Online]. Available: http://ieeexplore.ieee.org/document/5767843/

[185] P. Godfrey, J. Gryz, P. Lasek, and N. Razavi, "Interactive visualization of big data," in *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*. 2016, pp. 3–22. [Online]. Available: http://link.springer.com/10.1007/978-3-319-34099-9%7B%5C_%7D1

[186] M. Sarwat, "Interactive and scalable exploration of big spatial data—A data management perspective," in *Proc. 16th IEEE Int. Conf. Mobile Data Manage.*, vol. 1, Jun. 2015, pp. 263–270. [Online]. Available: http://ieeexplore.ieee.org/document/7264331/

[187] R. Neamtu, R. Ahsan, E. Rundensteiner, and G. Sarkozy, "Interactive time series exploration powered by the marriage of similarity distances," *Proc. VLDB Endowment*, vol. 10, no. 3, pp. 169–180, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3021924.3021933

[188] F. Bugiotti, D. Bursztyn, A. Deutsch, and I. Ileana, "Invisible glue: Scalable self-tuning multi-stores," in *Proc. CIDR*, 2015. [Online]. Available: http://cidrdb.org/cidr2015/Papers/CIDR15

[189] A. Abouzied, D. J. Abadi, and A. Silberschatz, "Invisible loading," in *Proc. 16th Int. Conf. Extending Database Technol. (EDBT)*, 2013, pp. 1–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=2452376.2452377

[190] J. X. Yu, L. Qin, and L. Chang. (2010). *Keyword Search in Relational Databases: A Survey*. [Online]. Available: https://www.researchgate.net/profile/Lijun

[191] S. Agarwal *et al.*, "Knowing when you're wrong: Building fast and reliable approximate query processing systems," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 481–492. [Online]. Available: http://dl.acm.org/citation.cfm?id=2593667

[192] S. Liu, B. Song, S. Gangam, L. Lo, and K. Elmeleegy, "Kodiak: leveraging materialized views for very low-latency analytics over high-dimensional web-scale data," *Proc. VLDB Endowment*, vol. 9, no. 13, pp. 1269–1280, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3007263.3007266
[193] M. Sadoghi, S. Bhattacherjee, B. Bhattacharjee, and M. Canim. (2016). "L-store: A real-time OLTP and OLAP system." [Online]. Available: https://arxiv.org/abs/1601.04084
[194] A. Bonifati, R. Ciucanu, and A. Lemay, "Learning path queries on graph databases," in *Proc. Int. Conf. Extending Database Technol. (EDBT)*, 2015, pp. 109–120. [Online]. Available: https://openproceedings.org/2015/conf/edbt/paper-6.pdf
[195] D. Deutch and A. Gilad. (2016). *Learning Queries from Examples and Their Explanations*. [Online]. Available: https://pdfs.semanticscholar.org/929d/e87b7eef6cac5609d607c576f44ae43adb56.pdf
[196] A. Abouzied, D. Angluin, C. Papadimitriou, J. M. Hellerstein, and A. Silberschatz, "Learning and verifying quantified Boolean queries by example," in *Proc. 32nd ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, 2013, pp. 49–60. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2463664.2465220
[197] L. Zhang, Y. Wang, and X. Xu, "Logic-partition based gaussian sampling for online aggregation," in *Proc. 5th Int. Conf. Adv. Cloud Big Data (CBD)*, Aug. 2017, pp. 182–187. [Online]. Available: http://ieeexplore.ieee.org/document/8026934/
[198] V. Alvarez, F. M. Schuhknecht, J. Dittrich, and S. Richter, "Main memory adaptive indexing for multi-core systems," in *Proc. 10th Int. Workshop Data Manage. New Hardw.*, 2014, p. 3, [Online]. Available: https://dl.acm.org/citation.cfm?id=2619231
[199] G. Reeves, J. Liu, S. Nath, and F. Zhao, "Managing massive time series streams with multi-scale compressed trickles," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 97–108, 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1687639
[200] T. Sellam and M. L. Kersten, "Meet charles, big data query advisor," in *Proc. CIDR*, 2013. [Online]. Available: http://www.cidrdb.org/cidr2013/Papers/CIDR13
[201] Y. Denneulin, C. Labbé, L. d'Orazio, and C. Roncancio, "Merging file systems and data bases to fit the grid," in *Proc. Data Manage. Grid Peer-to-Peer Syst.*, vol. 6265. 2010, pp. 13–25. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-15108-8
[202] S. Idreos, S. Manegold, H. Kuno, and G. Graefe, "Merging what's cracked, cracking what's merged: Adaptive indexing in main-memory column-stores," *Proc. VLDB Endowment*, vol. 4, no. 9, pp. 586–597, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2002938.2002944
[203] B. Qarabaqi and M. Riedewald, "Merlin: Exploratory analysis with imprecise queries," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 2, pp. 342–355, Feb. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7312990/
[204] H. A. Khan and M. A. Sharaf, "Model-based diversification for sequential exploratory queries," *Data Sci. Eng.*, vol. 2, no. 2, pp. 151–168, 2017. [Online]. Available: https://link.springer.com/article/10.1007
[205] S. Garg, J. E. Nam, I. V. Ramakrishnan, and K. Mueller, "Model-driven visual analytics," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, Oct. 2008, pp. 19–26. [Online]. Available: http://ieeexplore.ieee.org/document/4677352/
[206] K. S. Perera, M. Hahmann, W. Lehner, T. B. Pedersen, and C. Thomsen, "Modeling large time series for efficient approximate query processing," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2015, pp. 190–204. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-22324-7%7B%5C_%7D16
[207] J. Chen, Y. Wang, J. Liu, and Y. Huang, "Modeling semantic and behavioral relations for query suggestion," in *Web-Age Information Management*. 2013, pp. 666–678. [Online]. Available: http://link.springer.com/10.1007/978-3-642-38562-9
[208] H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis, "MuVE: Efficient multi-objective view recommendation for visual data exploration," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 731–742. [Online]. Available: http://ieeexplore.ieee.org/document/7498285/
[209] I. Alagiannis, R. Borovica, M. Branco, S. Idreos, and A. Ailamaki, "NoDB: efficient query execution on raw data files," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2012, vol. 58, no. 12, pp. 112–121. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2847579.2830508
[210] A. Albarrak, T. Noboa, H. A. Khan, M. A. Sharaf, X. Zhou, and S. Sadiq, "ORange: Objective-aware range query refinement," in *Proc. IEEE 15th Int. Conf. Mobile Data Manage.*, vol. 1, Jul. 2014, pp. 333–336. [Online]. Available: http://ieeexplore.ieee.org/document/6916939/
[211] P. Terlecki, F. Xu, M. Shaw, V. Kim, and R. Wesley, "On improving user response times in tableau," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2015, pp. 1695–1706. [Online]. Available: http://doi.acm.org/10.1145/2723372.2742799
[212] F. Bonchi, F. Giannotti, C. Lucchese, S. Orlando, R. Perego, and R. Trasarti, "On interactive pattern mining from relational databases," in *Knowledge Discovery in Inductive Databases*, vol. 4747. 2007, pp. 42–62. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-75549-4
[213] M. R. Vieira *et al.*, "On query result diversification," in *Proc. IEEE 27th Int. Conf. Data Eng.*, Apr. 2011, pp. 1163–1174. [Online]. Available: http://ieeexplore.ieee.org/document/5767846/
[214] A. Siddiqa, A. Karim, T. Saba, and V. Chang, "On the analysis of big data indexing execution strategies," *J. Intell. Fuzzy Syst.*, vol. 32, no. 5, pp. 3259–3271, 2017. [Online]. Available: https://eprints.soton.ac.uk/399946/
[215] R. Soulé and B. Gedik. (2014). "Optimized disk layouts for adaptive storage of interaction graphs." [Online]. Available: https://arxiv.org/abs/1410.5290
[216] M. M. M. Fuad, "Optimized multi-resolution indexing and retrieval scheme of time series," in *Progress in Artificial Intelligence*. 2015, pp. 603–608. [Online]. Available: http://link.springer.com/10.1007/978-3-319-23485-4
[217] K. T. Sridhar and M. A. Sakkeer, "Optimizing database load and extract for big data era," in *Database Systems for Advanced Applications*, vol. 8422. 2014, pp. 503–512. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-05813-9
[218] H. V. Jagadish, A. Nandi, and L. Qian, "Organic databases," in *Databases in Networked Information Systems*, vol. 7108. 2011, pp. 49–63. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-25731-5
[219] S. Wu, G. Chen, X. Zhou, Z. Zhang, A. K. H. Tung, and M. Winslett, "PABIRS: A data access middleware for distributed file systems," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 113–124. [Online]. Available: http://ieeexplore.ieee.org/document/7113277/
[220] D. Alabi and E. Wu, "PFunk-H: approximate query processing using perceptual models," in *Proc. Workshop Hum.-Loop Data Anal.*, 2016, p. 10. [Online]. Available: https://dl.acm.org/citation.cfm?id=2939512
[221] S. Idreos *et al.*, "Past and future steps for adaptive storage data systems: From shallow to deep adaptivity," in *Proc. Int. Workshop Enabling Real-Time Bus. Intell.*, 2016. [Online]. Available: https://stratos.seas.harvard.edu/publications/past-and-future-steps-adaptive-storage-data-systems-shallow-deep-adaptivity
[222] H. A. Khan and M. A. Sharaf, "Progressive diversification for column-based data exploration platforms," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 327–338. [Online]. Available: http://ieeexplore.ieee.org/document/7113295/
[223] D. Deutch and A. Gilad, "QPlain: Query by explanation," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 1358–1361. [Online]. Available: http://ieeexplore.ieee.org/document/7498344/?section=abstract
[224] M. Eirinaki and S. Patel, "QueRIE reloaded: Using matrix factorization to improve database query recommendations," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct./Nov. 2015, pp. 1500–1508. [Online]. Available: http://ieeexplore.ieee.org/document/7363913/
[225] V. Kantere, "Query similarity for approximate query answering," in *Database and Expert Systems Applications*, vol. 9828, S. Hartmann and H. Ma, Eds. Springer. 2016, pp. 355–367. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-44406-2
[226] K. Zoumpatianos, Y. Lou, T. Palpanas, and J. Gehrke, "Query workloads for data series indexes," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1603–1612. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2783258.2783382
[227] Q. T. Tran, C.-Y. Chan, and S. Parthasarathy, "Query by output," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2009, pp. 535–548. [Online]. Available: https://dl.acm.org/citation.cfm?doid=1559845.1559902
[228] H. Li, C.-Y. Chan, and D. Maier, "Query from examples: An iterative, data-driven approach to query construction," *Proc. VLDB Endowment*, vol. 8, no. 13, pp. 2158–2169, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2831369

[229] W. Fan, F. Geerts, Y. Cao, T. Deng, and P. Lu, "Querying big data by accessing small data," in *Proc. 34th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst. (PODS)*, 2015, pp. 173–184. [Online]. Available: http://dl.acm.org/citation.cfm?id=2745754.2745771

[230] P. Meisen, D. Keng, T. Meisen, M. Recchioni, and S. Jeschke, "Querying time interval data," in *Enterprise Information Systems*, 2015, pp. 45–68. [Online]. Available: http://link.springer.com/10.1007/978-3-319-29133-8

[231] A. Thiagarajan and S. Madden, "Querying continuous functions in a database system," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2008, pp. 791–804. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1376616.1376696

[232] S. Kandula *et al.*, "Quickr: Lazily approximating complex adhoc queries in bigdata clusters," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2016, pp. 631–646, doi: 10.1145/2882903.2882940.

[233] Q. Chen, M. Hsu, R. Wu, and J. Shan, "R-proxy framework for in-DB data-parallel analytics," in *Database and Expert Systems Applications* (Lecture Notes in Computer Science), vol. 7447. Berlin, Germany: Springer, 2012, pp. 266–280. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-32597-7%7B%5C_%7D24

[234] H.-T. Zheng and Y.-C. Zhang, "RDQS: A relevant and diverse query suggestion generation framework," in *Web Technologies and Applications*. Cham, Switzerland: Springer, 2015, pp. 586–597. [Online]. Available: http://link.springer.com/10.1007/978-3-319-25255-1%7B%5C_%7D48

[235] X. Ge, Y. Xue, Z. Luo, M. A. Sharaf, and P. K. Chrysanthis, "REQUEST: A scalable framework for interactive construction of exploratory queries," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 646–655. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7840657/

[236] R. Soulé and B. Gedik, "RailwayDB: Adaptive storage of interaction graphs," *VLDB J.*, vol. 25, no. 2, pp. 151–169, Apr. 2016, doi: 10.1007/978-1-4612-4380-9_41.

[237] A. Kim, E. Blais, A. Parameswaran, P. Indyk, S. Madden, and R. Rubinfeld, "Rapid sampling for visualizations with ordering guarantees," *Proc. VLDB Endowment*, vol. 8, no. 5, pp. 521–532, 2015.

[238] V. Silva *et al.*, "Raw data queries during data-intensive parallel workflow execution," *Future Gener. Comput. Syst.*, vol. 75, pp. 402–422, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17300237

[239] D. Basu *et al.*, "Regularized cost-model oblivious database tuning with reinforcement learning," in *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXVIII*, A. Hameurlain, J. Küng, R. Wagner, and Q. Chen, Eds. Berlin, Germany: Springer, 2016, pp. 96–132. [Online]. Available: https://link.springer.com/chapter/10.1007%7B%5C%%7D2F978-3-662-53455-7%7B%5C_%7D5

[240] Y. Hu, W. Zhai, Y. Tian, and T. Gao, "Research and application of query rewriting based on materialized views," in *Information and Automation*, vol. 86. Berlin, Germany: Springer, 2011, pp. 85–91. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-19853-3%7B%5C_%7D12

[241] R. F. Munir, O. Romero, A. Abelló, B. Bilalli, M. Thiele, and W. Lehner, "Resilient store: A heuristic-based data format selector for intermediate results," in *Model and Data Engineering* (Lecture Notes in Computer Science), vol. 9893. Cham, Switzerland: Springer, 2016, pp. 42–56. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-45547-1%7B%5C_%7D4

[242] A. Galakatos, A. Crotty, E. Zgraggen, C. Binnig, and T. Kraska, "Revisiting reuse for approximate query processing," *Proc. VLDB Endowment*, vol. 10, no. 1, pp. 1142–1153, Jun. 2017. [Online]. Available: http://www.vldb.org/pvldb/vol10/p1142-galakatos.pdf

[243] F. Psallidas, B. Ding, K. Chakrabarti, and S. Chaudhuri, "S4: Top-k spreadsheet-style search for query discovery," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2015, pp. 2001–2016, doi: 10.1145/2723372.2749452.

[244] Y. Cheng and F. Rusu, "SCANRAW: A database meta-operator for parallel *in-situ* processing and loading," *ACM Trans. Database Syst.*, vol. 40, no. 3, pp. 1–45, Oct. 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2838914.2818181

[245] F. Tauheed, T. Heinis, F. Schürmann, H. Markram, and A. Ailamaki, "SCOUT: Prefetching for latent structure following queries," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1531–1542, Jul. 2012, doi: 10.14778/2350229.2350267.

[246] S. S. Husain, A. Kalinin, A. Truong, and I. D. Dinov, "SOCR data dashboard: An integrated big data archive mashing medicare, labor, census and econometric information," *J. Big Data*, vol. 2, no. 1, p. 13, 2015. [Online]. Available: http://www.journalofbigdata.com/content/2/1/13

[247] R. Christensen, L. Wang, F. Li, K. Yi, J. Tang, and N. Villa, "STORM: Spatio-temporal online reasoning and management of large spatio-temporal data," *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1111–1116. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2723372.2735373

[248] B. Ding, S. Huang, S. Chaudhuri, K. Chakrabarti, and C. Wang, "Sample + seek: Approximating aggregates with distribution precision guarantee," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 679–694. [Online]. Available: https://dl.acm.org/citation.cfm?id=2915249

[249] B. C. Kwon, J. Verma, P. J. Haas, and Ç. Demiralp, "Sampling for scalable visual analytics," *IEEE Comput. Graph. Appl.*, vol. 37, no. 1, pp. 100–108, Jan./Feb. 2017. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7819391/

[250] I. Psaroudakis *et al.*, "Scaling up mixed workloads: A battle of data freshness, flexibility, and scheduling," in *Performance Characterization and Benchmarking. Traditional to Big Data*, R. Nambiar and M. Poess, Eds. Cham, Switzerland: Springer, 2015, pp. 97–112. [Online]. Available: https://link.springer.com/chapter/10.1007%7B%5C%%7D2F978-3-319-15350-6%7B%5C_%7D7

[251] A. W.-C. Fu, E. Keogh, L. Y. H. Lau, C. A. Ratanamahatana, and R. C.-W. Wong, "Scaling and time warping in time series querying," *VLDB J.*, vol. 17, no. 4, pp. 899–921, 2008. [Online]. Available: http://link.springer.com/article/10.1007/s00778-006-0040-z

[252] L. Sidirourgos, M. Kersten, and P. Boncz, "SciBORQ: Scientific data management with bounds on runtime and quality," *Proc. CIDR*, 2011, pp. 296–301.

[253] L. Sidirourgos, M. Kersten, and P. Boncz, "Scientific discovery through weighted sampling," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 300–306. [Online]. Available: http://ieeexplore.ieee.org/document/6691587/

[254] A. Kalinin and S. Zdonik, "Searchlight: Enabling integrated search and exploration over large multidimensional data," *Proc. VLDB Endowment*, vol. 8, no. 10, pp. 1094–1105, 2015. [Online]. Available: http://www.vldb.org/pvldb/vol8/p1094-kalinin.pdf

[255] A. Parameswaran, N. Polyzotis, and H. Garcia-Molina, "SeeDB: Visualizing database queries efficiently," *Proc. VLDB Endowment*, vol. 7, no. 4, pp. 325–328, Dec. 2013, doi: 10.14778/2732240.2732250.

[256] A. Pavlo *et al.*, "Self-driving database management systems," in *Proc. CIDR*, 2017, pp. 1–6. [Online]. Available: http://db.cs.cmu.edu/papers/2017/p42-pavlo-cidr17.pdf

[257] S. Idreos, M. L. Kersten, and S. Manegold, "Self-organizing tuple reconstruction in column-stores," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2009, pp. 297–308. [Online]. Available: https://dl.acm.org/citation.cfm?doid=1559845.1559878

[258] T. Sellam, E. Müller, and M. Kersten, "Semi-automated exploration of data warehouses," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2015, pp. 1321–1330, doi: 10.1145/2806416.2806538.

[259] L. Sun, M. J. Franklin, J. Wang, and E. Wu, "Skipping-oriented partitioning for columnar layouts," *Proc. VLDB Endowment*, vol. 10, no. 4, pp. 421–432, 2016-11, doi: 10.14778/3025111.3025123.

[260] M. Olma, M. Karpathiotakis, I. Alagiannis, M. Athanassoulis, and A. Ailamaki, "Slalom: Coasting through raw data via adaptive partitioning and indexing," *Proc. VLDB Endowment*, vol. 10, no. 10, pp. 1106–1117, 2017. [Online]. Available: https://dl.acm.org/citation.cfm?id=3115415

[261] L. Jiang and A. Nandi, "SnapToQuery: Providing interactive feedback during exploratory query specification," *Proc. VLDB Endowment*, vol. 8, no. 11, pp. 1250–1261, 2015.

[262] L. Wang, R. Christensen, F. Li, and K. Yi, "Spatial online sampling and aggregation," *Proc. VLDB Endowment*, vol. 9, no. 3, pp. 84–95, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2850583.2850584

[263] F. Lei *et al.*, "Speed up distance-based similarity query using multiple threads," in *Proc. 6th Int. Symp. Parallel Archit., Algorithms Program.*, Jul. 2014, pp. 215–219. [Online]. Available: http://ieeexplore.ieee.org/document/6916467/?reload=true

[264] S. Krishnan, J. Wang, M. J. Franklin, K. Goldberg, and T. Kraska, "Stale view cleaning: Getting fresh answers from stale materialized views," *Proc. VLDB Endowment*, vol. 8, no. 12, pp. 1370–1381, Aug. 2015, doi: 10.14778/2824032.2824037.

[265] F. Halim, S. Idreos, P. Karras, and R. H. C. Yap, "Stochastic database cracking: Towards robust adaptive indexing in main-memory column-stores," *Proc. VLDB Endowment*, vol. 5, no. 6, pp. 502–513, 2012. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2168651.2168652

[266] Y. Wang, L. Chen, and G. Agrawal, "Supporting online analytics with user-defined estimation and early termination in a MapReduce-like framework," in *Proc. ACM Int. Workshop Data-Intensive Scalable Comput. Syst. (DISCS)*, 2015, pp. 1–8. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2831244.2831247

[267] A. Bondu, M. Boullé, and A. Cornuéjols, "Symbolic representation of time series: A hierarchical coclustering formalization," in *Advanced Analysis and Learning on Temporal Data*, vol. 9785. Cham, Switzerland: Springer, 2016, pp. 3–16. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-44412-3%7B%5C_%7D1

[268] G. Cormode, "Synopses for massive data: Samples, histograms, wavelets, sketches," *Found. Trends Databases*, vol. 4, nos. 1–3, pp. 1–294, 2011. [Online]. Available: http://www.nowpublishers.com/article/Details/DBS-004

[269] K. Zeng, S. Gao, B. Mozafari, and C. Zaniolo, "The analytical bootstrap: A new method for fast error estimation in approximate query processing," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2014, pp. 277–288, doi: 10.1145/2588555.2588579.

[270] E. Wu, L. Battle, and S. R. Madden, "The case for data visualization management systems: vision paper," *VLDB J.*, vol. 7, no. 10, pp. 903–906, 2012.

[271] P. Cudre-Mauroux, E. Wu, and S. Madden, "The case for rodentstore, an adaptive, declarative storage system," in *Proc. CIDR*, 2009. [Online]. Available: http://www-db.cs.wisc.edu/cidr/cidr2009/Paper%7B%5C_%7D26.pdf%7B%5C%%7D5Cnpapers2://publication/uuid/58452AC7-980C-496A-B266-E73DA96094FF

[272] M. L. Kersten, S. Idreos, S. Manegold, and E. Liarou, "The researcher's guide to the data deluge: Querying a scientific database in just a few seconds," in *Proc. VLDB Endowment*, 2011, p. 1474.

[273] P. Baumann, A. M. Dumitru, and V. Merticariu, "The array database that is not a database: File based array query answering in rasdaman," in *Advances in Spatial and Temporal Databases* (Lecture Notes in Computer Science), vol. 8098. Berlin, Germany: Springer, 2013, pp. 478–483. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-40235-7%7B%5C_%7D32

[274] D. Tsoumakos and C. Mantas, "The case for multi-engine data analytics," in *Euro-Par 2013: Parallel Processing Workshops* (Lecture Notes in Computer Science), vol. 8374. Berlin, Germany: Springer, 2014, pp. 406–415. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-54420-0%7B%5C_%7D40

[275] G. Chernishev, "The design of an adaptive column-store system," *J. Big Data*, vol. 4, no. 1, p. 5, Dec. 2017, doi: 10.1186/s40537-017-0069-4.

[276] S. Venkataraman *et al.*, "The power of choice in data-aware cluster scheduling," in *Proc. USENIX Conf. Oper. Syst. Design Implement.*, 2014, pp. 301–316.

[277] F. M. Schuhknecht, A. Jindal, and J. Dittrich, "The uncracked pieces in database cracking," *Proc. VLDB Endowment*, vol. 7, no. 2, pp. 97–108, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2732228.2732229

[278] T. S. Nguyen and T. A. Duong, "Time series subsequence matching based on a combination of PIP and clipping," in *Intelligent Information and Database Systems* (Lecture Notes in Computer Science), vol. 6591. Berlin, Germany: Springer, 2011, pp. 149–158. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-20039-7%7B%5C_%7D15

[279] T. N. Dang and L. Wilkinson, "TimeExplorer: Similarity search time series by their signatures," in *Advances in Visual Computing* (Lecture Notes in Computer Science), vol. 8033. Berlin, Germany: Springer, 2013, pp. 280–289. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-41914-0%7B%5C_%7D28

[280] D. A. Aoyama, J.-T. T. Hsiao, A. F. Cárdenas, and R. K. Pon, "TimeLine and visualization of multiple-data sets and the visualization querying challenge," *J. Vis. Lang. Comput.*, vol. 18, no. 1, pp. 1–21, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1045926X06000395

[281] K. Feng, G. Cong, S. S. Bhowmick, W.-C. Peng, and C. Miao, "Towards best region search for data exploration," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 1055–1070. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2882903.2882960

[282] J. Dittrich and A. Jens, "Towards a one size fits all database architecture," in *Proc. CIDR*, 2011, pp. 195–198.

[283] T. Eavis and A. Taleb, "Towards a scalable, performance-oriented OLAP storage engine," in *Database Systems for Advanced Applications* (Lecture Notes in Computer Science), vol. 7239. Berlin, Germany: Springer, 2012, pp. 185–202. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-29035-0%7B%5C_%7D13

[284] A. M. Khan, D. Gonçalves, and D. C. Leão, "Towards an adaptive framework for real-time visualization of streaming big data," in *EuroVis 2017-Posters*, A. P. Puig and T. Isenberg, Eds. Eurographics Association, 2017, pp. 13–15. [Online]. Available: https://diglib.eg.org/handle/10.2312/eurp20171155

[285] O. O. Akande and P. J. Rhodes, "Towards an efficient storage and retrieval mechanism for large unstructured grids," *Future Gener. Comput. Syst.*, vol. 45, pp. 53–69, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X14002180

[286] S. Richter, J.-A. Q. Ruiz, S. Schuh, and J. Dittrich, "Towards zero-overhead static and adaptive indexing in Hadoop," *VLDB J.*, vol. 23, no. 3, pp. 469–494, 2014.

[287] D. Moritz, D. Fisher, B. Ding, and C. Wang, "Trust, but verify: Optimistic visualizations of approximate queries for exploring big data," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 2904–2915. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3025453.3025456

[288] L. Deri, S. Mainardi, and F. Fusco, "Tsdb: A compressed database for time series," in *Traffic Monitoring and Analysis* (Lecture Notes in Computer Science), vol. 7189. Berlin, Germany: Springer, 2012, pp. 143–156. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-28534-9%7B%5C_%7D16

[289] S. Idreos, M. L. Kersten, and S. Manegold, "Updating a cracked database," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2007, pp. 413–413. [Online]. Available: https://dl.acm.org/citation.cfm?doid=1247480.1247527

[290] M. Mayer *et al.*, "User interaction models for disambiguation in programming by example," in *Proc. 28th Annu. ACM Symp. User Interface Softw. Technol.*, 2015, pp. 291–301. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2807442.2807459

[291] A. Dhankar and V. Singh, "User search intention in interactive data exploration: A brief review," in *Advances in Computing and Data Sciences*, vol. 721, M. Singh, P. K. Gupta, V. Tyagi, A. Sharma, T. Ören, and W. Grosky, Eds. Singapore: Springer, 2017, pp. 409–419, doi: 10.1007/978-981-10-5427-3_44.

[292] A. Al-Naser, M. Rasheed, D. Irving, and J. Brooke, "User's interpretations of features in visualization," in *Computer Vision, Imaging and Computer Graphics—Theory and Applications*, vol. 550. 2015, pp. 97–114. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-25117-2%7B%5C_%7D7

[293] B. Qarabaqi and M. Riedewald, "User-driven refinement of imprecise queries," in *Proc. IEEE 30th Int. Conf. Data Eng. Workshops*, Mar./Apr. 2014, pp. 916–927. [Online]. Available: http://ieeexplore.ieee.org/document/6818355/

[294] P. Carvalho, P. Hitzelberger, B. Otjacques, F. Bouali, and G. Venturini, "Using information visualization to support open data integration," in *Data Management Technologies and Applications*. Cham, Switzerland: Springer, 2015, pp. 1–15. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-25936-9%7B%5C_%7D1

[295] U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl, "VDDA: Automatic visualization-driven data aggregation in relational databases," *VLDB J.*, vol. 25, no. 1, pp. 53–77, 2016-02, doi: 10.1007/s00778-015-0396-z.

[296] W. Zhao, Y. Cheng, and F. Rusu, "Vertical partitioning for query processing over raw data," in *Proc. ACM 27th Int. Conf. Sci. Stat. Database Manage. (SSDBM)*, 2015, pp. 1–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=2791347.2791369

[297] M. Komenda and D. Schwarz, "Visual analytics in environmental research: A survey on challenges, methods and available tools," in *Environmental Software Systems. Fostering Information Sharing*, vol. 413. Berlin, Germany: Springer, 2013, pp. 618–629. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-41151-9%7B%5C_%7D58

[298] J. Sjöbergh and Y. Tanaka, "Visual data exploration using webbles," in *Webble Technology*, vol. 372. Berlin, Germany: Springer, 2013, pp. 119–128, doi: 10.1007/978-3-642-38836-1_10.

[299] M. Kahng, D. Fang, and D. H. P. Chau, "Visual exploration of machine learning results using data cube analysis," in *Proc. Workshop Hum.-Loop Data Anal.*, 2016, pp. 1–6. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2939502.2939503

[300] A. Malagoli *et al.*, "Visual query specification and interaction with industrial engineering data," in *Advances in Visual Computing* (Lecture Notes in Computer Science), vol. 8034. Berlin, Germany: Springer, 2013, pp. 58–67. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-41939-3%7B%5C_%7D6

[301] H. Cao, Y. Li, C. M. Allen, M. A. Phinney, and C.-R. Shyu, "Visual reasoning indexing and retrieval using in-memory computing," *Int. J. Semantic Comput.*, vol. 10, no. 3, pp. 17–24, 2016. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S1793351X16400110

[302] Y. Park, M. Cafarella, and B. Mozafari, "Visualization-aware sampling for very large databases," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 755–766. [Online]. Available: https://arxiv.org/abs/1510.03921%20http://ieeexplore.ieee.org/document/7498287/

[303] E. Olshannikova, A. Ometov, Y. Koucheryavy, and T. Olsson, "Visualizing big data with augmented and virtual reality: Challenges and research agenda," *J. Big Data*, vol. 2, no. 1, p. 22, 2015. [Online]. Available: http://www.journalofbigdata.com/content/2/1/22

[304] C. Combi and B. Oliboni, "Visually defining and querying consistent multi-granular clinical temporal abstractions," *Artif. Intell. Med.*, vol. 54, no. 2, pp. 75–101, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365711001424

[305] A. Key, B. Howe, D. Perry, and C. R. Aragon, "VizDeck: Self-organizing dashboards for visual analytics," in *Proc. 2nd Workshop Hum.-Loop Data Anal.*, 2012, pp. 681–684. [Online]. Available: https://dl.acm.org/citation.cfm?id=2213931

[306] D. Moritz and D. Fisher, "What users don't expect about exploratory data analysis on approximate query processing systems," in *Proc. 2nd ACM Workshop Hum.-Loop Data Anal. (HILDA)*, 2017, pp. 9-1–9-4, doi: 10.1145/3077257.3077258.

[307] H. Bian *et al.*, "Wide table layout optimization based on column ordering and duplication," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 299–314, doi: .10.1145/3035918.3035930

[308] F. Rusu, Z. Zhuang, M. Wu, and C. Jermaine, "Workload-driven antijoin cardinality estimation," *ACM Trans. Database Syst.*, vol. 40, no. 3, pp. 16-1–16-41, Oct. 2015. [Online]. Available: https://dl.acm.org/citation.cfm?id=2818178

[309] E. A. Rundensteiner *et al.*, "Xmdvtool^{Q}:: Quality-aware interactive data exploration," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2007, pp. 1109–1112, doi: .10.1145/1247480.1247623

[310] M. Drosou and E. Pitoura, "YmalDB: Exploring relational databases via result-driven recommendations," *VLDB J.*, vol. 22, no. 6, pp. 849–874, 2013. [Online]. Available: https://link.springer.com/article/10.1007/s00778-013-0311-4

[311] B. Wang, G. Chen, J. Bu, and Y. Yu, "ZoomTree: Unrestricted zoom paths in multiscale visual analysis of relational databases," in *Computer Vision, Imaging and Computer Graphics. Theory and Applications*, vol. 229. Berlin, Germany: Springer, 2011, pp. 299–317. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-25382-9

[312] S. Idreos and E. Liarou, "dbTouch: Analytics at your fingertips," in *Proc. CIDR*, 2013, pp. 1–11. [Online]. Available: http://oai.cwi.nl/oai/asset/21322/21322B.pdf

[313] K. Zeng, S. Agarwal, and I. Stoica, "iOLAP: Managing uncertainty for efficient incremental OLAP," in *Proc. SIGMOD ACM Int. Conf. Manage. Data*, 2016, pp. 1347–1361, doi: 10.1145/2882903.2915240.

ALEJANDRO ALVAREZ-AYLLON received the master's degree in computer science engineering from the University of Cádiz, in 2010, where he is currently pursuing the Ph.D. degree. He was with the European Organization for Nuclear Research, from 2009 to 2018, on different data management components for the LHC computing grid. He has been a Software Engineer with the Astronomy Department, University of Geneva, since 2018.

MANUEL PALOMO-DUARTE received the degree in computer science from the University of Seville and the Ph.D. degree from the University of Cádiz, Spain, where he is currently a Lecturer. His main research interests are creative computing, open data engineering, and collaboration, with a focus on the application of software technologies for computer-aided creation, exploration, and assessment, and he has published different contributions in indexed peer-reviewed journals and research conference proceedings in these fields.

JUAN-MANUEL DODERO received the Ph.D. degree in computer science engineering from the University Carlos III of Madrid. He is currently an Associate Professor with the University of Cádiz, a Senior Lecturer with the University Carlos III of Madrid, and also an ICT Research and Development Consultant in Spanish companies. His main research interests are creative computing and technology-enhanced learning, with a focus on the application of software technologies for computer-aided creation and assessment. He has participated in diverse research and development projects in relation to these subjects.